Yashwardhan Chaudhuri

Delhi, India | +91-9315525988 | yashwardhan20417@iiitd.ac.in | linkedin.com/in/yashwardhan-chaudhuri

Summary

Machine Learning Engineer with experience building production LLM, RAG, and agentic systems. Expertise in LangGraph, FastAPI, LoRA fine-tuning, hybrid retrieval (BM25+dense), vector databases (Qdrant), PostgreSQL, and LLM evaluation. Delivered AI systems across education, healthcare, and research workflows.

Professional Experience

Co-Founder (Product + Research) *Jan 2026 – Present*
Noteweave — evolved from ResXiv *Delhi, India*

- Designed **E3** evaluation protocol for automated paper critique (arXiv:2605.27072): **90.2%** partial-inclusive recall on **4,598** post-cutoff ICLR 2026 issue rows against GPT, Claude, and human review baselines.
- Shipped LangGraph agents with FastAPI serving to **150+** users on VS Code and web app, reducing time from research brief to cited, executable plans.
- Built hybrid search over **3M+ arXiv** papers (BM25 + dense, RRF) grounding agent and automated critique outputs.

Founder in Residence *Apr 2026 – May 2026*
Founders, Inc. — Canopy Accelerator (Noteweave) *San Francisco, CA*

- Canopy residency for Noteweave: shipped LangGraph agent pipelines, hybrid search, and E3 evaluation harnesses from research prototype to production release.

Co-Founder (Product + Tech) *Sep 2025 – Mar 2026*
ResXiv *Delhi, India*

- Shipped federated fusion search across **11 scholarly indexes** (**280M+** catalog records) with discipline tagging, deduplication, and on-demand PDF ingest into a shared project library.
- Reached **300+ MAU** on production beta serving R&D teams; highest engagement on fusion search, cited paper chat, and draft critique via FastAPI.
- Built end-to-end research OS backend: mandatory-citation RAG, multi-pass draft critique with OpenAlex checks, and LaTeX write workspace on PostgreSQL + Qdrant.

Independent Consultant (Part-Time) — ML & Search Systems *Jan 2026 – Mar 2026*
DraftNCraft (Global IP Services) *Remote*

- Enabled prior-art search over million-record patent corpora for Global IP Services: Elasticsearch + Qdrant with RRF and PostgreSQL metadata across USPTO/EPO/WIPO filings.
- Shortened multilingual patent ingestion through quantized RapidOCR + HunyuanMT 7B co-serving, warm loads, and OCR/translation caching on million-record batch runs.

Associate ML Scientist I *Oct 2023 – Sep 2025*
Wadhwani AI *New Delhi, India*

- **ORF / Vachan Samiksha:** contributed ASR scoring to statewide oral-reading fluency program in Gujarat/Rajasthan government schools; fine-tuned Wav2Vec2-style models with lexical–sublexical analysis on field transcripts (CER, WER, WCPM).
- **LEARN / SWAYAM:** delivered **10-course** national pilot for Ministry of Education SWAYAM — lecture segmentation (AVLectures to ASR + LLM production) and hybrid course search for national discovery.
- **PROS (Jhpiego):** tree-ensemble missed-visit risk model for antenatal care with SHAP interpretability; exported to Java via **m2cgen** for JDK edge deployment; field-validated with Jhpiego in Madhya Pradesh.

Research Experience

Research Associate *Jan 2022 – Jun 2024*
Usable Security Group @ IIIT-Delhi *New Delhi, India*

- Developed low-resource crowd-counting architectures with **80% FLOPs reduction** and **25%** benchmark improvement; co-authored FGA (IJCNN'24) including PyTorch training pipeline, ablations, and Grad-CAM interpretability analysis.
- Prototyped violence-detection dashboard on CCTV feeds for operator review: **0.01M-parameter** crowd-density regressor, VGG–LSTM temporal encoder, Streamlit UI.
- Benchmarked five pretrained audio encoders for deepfake detection; XLS-R achieved lowest equal-error rate on ASVspoof 2019 and In-the-Wild corpora.
- IJCNN 2024 technical program committee reviewer.

CHANAKYA Research Fellow *Sep 2022 – Mar 2023*
IIT Patna *Patna, Bihar*

– Engineered aerial crowd-density pipeline in PyTorch/OpenCV with ONNX export on NVIDIA Jetson Nano; adapted lightweight density-map regressors for flight-footage evaluation (INR **100,000** fellowship).

External Project

FedEm — distributed relay LoRA finetuning (Project Seshu)
Open Source AI Hackathon'24 — Hasgeek / Meta

– Released **fedem** on PyPI and GitHub (github.com/mlsquare/fedem) for distributed relay LoRA fine-tuning with adapter-lock scheduling and HuggingFace Hub integration.

Publications

*Google Scholar: h-index **3** — scholar.google.com/citations?user=VV2eC_gAAAAJ*

E3: Issue-Level Backtesting for Automated Research Critique *arXiv 2026*
Yashwardhan Chaudhuri, S Jain, P Mundra

– Issue-level backtesting framework for automated research critique with discrete entries encoding nature, location, and resolution path. Evaluated on **100 ICLR 2026 papers** (**4,598** issue rows) using post-cutoff corpus design and blinded meta-judge labels; **90.2%** partial-inclusive recall relative to GPT, Claude, and human review baselines. arxiv.org/abs/2605.27072

VoxMed: One-Step Respiratory Disease Classifier using Digital Stethoscope Sounds *INTERSPEECH'24*
Paridhi Mundra*, Manik Sharma*, **Yashwardhan Chaudhuri***, Orchid Chetia Phukan, Arun Balaji Buduru

– UI-assisted respiratory screening from stethoscope audio using an Audio Spectrogram Transformer feature extractor and 1-D CNN classifier head; evaluated on the ICBHI dataset (Greece and Portugal recordings). github.com/Sample-User131001/VoxMed

ASGIR: Audio Spectrogram Guided Classification And Information Retrieval For Birds *INTERSPEECH'24*
Yashwardhan Chaudhuri*, Paridhi Mundra*, Arnesh Batra*, Orchid Chetia Phukan, Arun Balaji Buduru

– Audio spectrogram-guided classification and information retrieval for bird vocalizations, coupled with location-conditioned Wikipedia retrieval for habitat context; evaluated on 51 Xeno-Canto species classes. github.com/MainSample1234/AS-GIR

FGA: Fourier-Guided Attention Network for Crowd Count Estimation *IJCNN'24*
Yashwardhan Chaudhuri, Ankit Kumar, Arun Balaji Buduru, Adel Alshamrani

– Fourier-guided attention module combining FFT-derived global features with convolutional local features for crowd density estimation; integrated into CSRNet and CANNet with improved MSE/MAE on ShanghaiTech-A/B, UCF-CC-50, and JHU++. arxiv.org/abs/2407.06110

Efficiency In Simplicity: A Lightweight Architecture for Resource-Efficient Crowd Counting *arXiv 2024*
Yashwardhan Chaudhuri, Ankit Kumar, Paridhi Mundra, Orchid Chetia Phukan, Arun Balaji Buduru

– AFFNet applies Adjacent Feature Fusion on lightweight backbones to aggregate multi-scale representations for crowd counting; reduces FLOPs versus prior architectures and improves estimation on ShanghaiTech-A/B and UCF-CC-50, with **13.31%** MSE reduction on ShanghaiTech-B. arxiv.org/abs/2401.05968

Education

Indraprastha Institute of Information Technology Delhi (IIIT-Delhi) *Jul 2024*
B.Tech, Computer Science with minors in Human Centered Design — CGPA: 8.0 *New Delhi, India*

Skills

– **Modeling:** PyTorch, HuggingFace, fine-tuning (Wav2Vec2, AST, CNN/LSTM), ablations, eval harnesses
– **LLMs & RAG:** LangGraph, LangChain, Ollama, LoRA fine-tuning, hybrid search (BM25 + dense, RRF), cited generation, LLM evaluation
– **ML deployment:** FastAPI, REST, WebSocket, PostgreSQL, MySQL, pgvector, AWS, Streamlit, quantization, Docker, GPU inference, ONNX
– **Domains:** ASR/speech, NLP, computer vision, audio classification, tabular ML (SHAP, m2cgen)