

Ashutosh Ranjan

authran.off@gmail.com | +91 8789901674 | Ravangla, Sikkim | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

Experience

ML Engineer Intern

June 2026 – Present

Navriti Technologies, Bangalore

- Built an end-to-end **text-to-video system** turning a **2,000-word article** into a narrated, captioned explainer video: **LangChain** + **vLLM** scene scripting, **22B LTX-2.3 diffusion model**, Kokoro TTS, **Whisper**-synced captions, orchestrated with **LangGraph**.
- Ran a **22B video transformer** + **12B text encoder** (~66 GB) on one **16 GB T4** via **4-bit NF4 quantization**, CPU offload, and VAE tiling (~5× VRAM cut); fixed **fp16-overflow** NaNs and $O(N^2)$ **attention** OOMs via measured memory budgeting.
- Engineered **ffmpeg** assembly (**Whisper**-timed karaoke captions, freeze-frame narration sync, 2× upscale) and a two-stage LLM chain distilling corpora into **Pydantic**-validated **What/Why/How/Where** scripts.

AI Engineer Intern

Dec 2025 – Feb 2026

Yantrapragya Developers Pvt. Ltd., Noida

- Trained NLP sentiment models for production feedback, improving **F1** by **8%** with custom preprocessing.

Software Development Intern

Jan 2025 – Mar 2025

AICTE — Ernst & Young GDS

- Improved frontend performance and SEO by **20%** through responsive design optimizations.
- Architected secure JWT-authenticated REST APIs serving **100,000+** daily user transactions.

Technical Skills

Languages: C, C++, Python, JavaScript, SQL

Generative AI & ML: PyTorch, TensorFlow, Hugging Face (Transformers, Diffusers), LangChain, LangGraph, vLLM, Diffusion Models, LLM Pretraining & Fine-Tuning, RLHF/**DPO**, Quantization (bitsandbytes), Distributed Training (**DDP**, **FSDP**), Mixed Precision, Whisper

Data Science: Pandas, NumPy, Scikit-learn, Natural Language Processing (NLP), Matplotlib, Streamlit

Web & Backend: React, Node.js, Express.js, REST APIs, JWT, MongoDB, FastAPI, Flask, Tailwind CSS

Infrastructure: Kubernetes, Docker, PostgreSQL, MinIO/S3, Prometheus, Grafana, Git, Linux, CI/CD, FFmpeg

Projects

AshuGPT — Decoder-Only LLM Pretrained From Scratch (PyTorch) | [Code](#) | [Live](#)

Jun 2026 – Aug 2026

- Pretrained a **124M-parameter** transformer from scratch on **2.46B tokens** of FineWeb-Edu to **23.53** validation perplexity — 20,000 steps, 27 h on one RTX 2080 Ti at **25K tokens/s** — building **RoPE**, **RMSNorm**, **SwiGLU**, KV-cache attention and byte-level **BPE** in **6,000 lines** of raw PyTorch, no transformers.
- Cut peak memory **50.9%** with gradient checkpointing, accumulation and fused attention, all **gradient-identical**; **fp16 DDP** converges **bit-identically** across shards, and **FSDP** + CPU offload shards a **1.23B** config to **2.04 GB/GPU**.
- Reclaimed the **89%** of each fine-tuning step lost to padding via **sequence packing** (block-diagonal masking, per-example RoPE positions, unit-tested gradient-identical) for **4.40x** supervised throughput; post-trained on Alpaca/Dolly (stop rate **30% to 100%**), **20,000 UltraChat** conversations, and **DPO** on HH-RLHF (**50% to 57.1%** win rate).
- Shipped both checkpoints to **Hugging Face Spaces** behind a streaming **FastAPI** service (KV-cached, 80 tok/s) under **441 tests** on **GitHub Actions**; sweeping a **2x2 factorial** on whether packing shifts the optimal learning rate.

AshML — Kubernetes-Native ML Training and Serving Platform | [Code](#) | [Live](#) | [HF Space](#)

Aug 2026

- Built a **Node.js** control plane that schedules training onto **Kubernetes**: its own scheduler enforcing quota and placement (and printing every node it rejected), an executor driving Jobs from observed Pod status, and **PostgreSQL** state behind an append-only event log and a **SKIP LOCKED** queue.
- Checkpoints upload to **MinIO/S3** presigned and are **verified against the bucket** before a model version may use them; one **PRODUCTION** version per model, promoted transactionally, then served behind a weighted **canary router**.
- ResNet-18/CIFAR-10** ran through all of it: one epoch, 691 s, **65.59%** top-1 re-evaluated from object store, **re-producible** across reruns; **SIGKILL** the training pod and it resumes weights, optimizer, LR schedule *and* data order.

AshCode — Terminal Coding Agent (Published on PyPI) | [Code](#) | [PyPI](#) | [Demo](#)

Jul 2026

- Engineered a streaming agentic loop exposing **12 tools** (file edit/patch with diffs, shell with timeout, scrubbed env, grep/glob, web search) plus **MCP** servers over stdio and HTTP/SSE, on any OpenAI-compatible endpoint.
- Sustained long sessions in a finite context window by compacting history at **80%** capacity and pruning stale tool output, with a signature-based detector that breaks repeating failed tool calls; **published to PyPI** under **36 tests** on **GitHub Actions** across **Python 3.11–3.13**, with **6 approval policies** gating destructive operations.

Education

National Institute of Technology Sikkim

Aug 2023 – May 2027

B.Tech in Computer Science and Engineering

CGPA: **7.5**

Kendriya Vidyalaya NHPC, Singtam

Class XII: **95%** | Class X: **96.6%**

Achievements

- Competitive programmer with a **1900 rating (Knight)** on [LeetCode](#), having solved **320+** DSA problems.
- Selected for **Amazon ML Summer School 2026** — top 3,000 of 150,000+ applicants (approx. 2%).
- Secured **6th place** in the regional **Smart India Hackathon** with an AI-powered [Garbage Segregation model](#).