Atandra Bharati

Deep Learning Research Engineer · From-scratch PyTorch implementations of frontier AI architectures

atandrabharati@gmail.com github.com/atandra2000
linkedin.com/in/atandrabharati atandra2000.github.io/mycv
Kolkata, India · Remote-friendly · Open to global remote roles

SUMMARY

Self-taught deep learning research engineer (B.Tech Civil Engineering, 2024) with 14 from-scratch PyTorch projects spanning LLMs, latent diffusion, multimodal AI, agentic research, long-context attention, and state-space models. Headline results: 78% peak-memory cut on LLaMA-3 pretraining (92 GB 20 GB on a single A100 80GB); 0.0947 training loss on Stable Diffusion 1.x from scratch (860M UNet, 2× RTX 5090); 2× KV-cache reduction at 128K in a GPT-OSS-style long-context MoE; 15-phase multi-agent ML research platform with 878 passing tests.

TECHNICAL SKILLS

LLM LLaMA-3 · DeepSeek-V3 · GPT-OSS · Mamba-3 · MLA · GQA · MoE · MTP · Gated Delta Net · complex64 SSD · sliding/full attention · learned sinks · YaRN · μP · WSD · NorMuon · CautiousAdamW

Generative Vision Stable Diffusion 1.x · latent diffusion · UNet · DDPM/DDIM · Min-SNR · CFG · EMA · VAE · GAN · DCGAN · PatchGAN · β-VAE · CycleGAN · AdaIN

Multimodal & Video ViT · SigLIP · PaliGemma-style fusion · HRNet pose · ST-GCN · CTR-GCN · NTU RGB+D 120

Core Python 3.12 · PyTorch 2.x · torch.compile · SDPA · Flash-Attn 2 · BF16 · chunked CE · gradient checkpointing · DDP · safetensors · HuggingFace · Diffusers · W&B · Comet

Agentic & Infra Multi-agent orchestration · provider-agnostic LLM routing · vector + graph memory · Pydantic v2 · pytest · Ruff · A100 80GB · RTX 5090 · RTX 6000 Ada · RTX 3090 · P100 · 2× T4

EXPERIENCE

ML Engineering Portfolio Nov 2022 – Present · Self-directed · github.com/atandra2000

- **14 from-scratch PyTorch systems** across LLMs, latent diffusion, multimodal AI, generative vision, video understanding, and agentic ML — no HF Trainer, no Lightning, every layer written by hand.
- **Memory engineering flagship: LLaMA-3-Lite** 515M-param LLaMA-3-style transformer cut peak training memory 92 GB 20 GB (78%) on a single A100 80GB via gradient checkpointing, chunked cross-entropy (logits 50 GB 0.3 GB), disk-backed token caching (RAM 112 GB 1 MB), BF16, FA2, channels_last, and fused AdamW.
- **Frontier reproductions** faithful DeepSeek-V3 (MLA + AuxLossFreeGate MoE + MTP, with the absorption trick), GPT-OSS (sliding/full attention alt + learned sinks + YaRN 128K, 2× KV-cache cut at 128K), and Mamba-3 (complex64 SSD with 50% smaller state, MIMO head mixing, zero causal conv).
- **Diffusion flagship: Stable Diffusion 1.x** 860M-param UNet trained from random init on 2× RTX 5090 across a 7-phase curriculum (LAION-Aesthetic DiffusionDB/JourneyDB VGGFace2 COCO consolidation); best loss 0.0947 at epoch 16; epoch-42 checkpoint released on HuggingFace.
- **Agentic flagship: Autonomous ML Research Engineer** 15-phase multi-agent platform (23 agents, 61 tools, 186 models, 878 tests) that turns an arXiv paper into evaluated experiments end-to-end with provider-agnostic LLM routing and self-repair.

SELECTED PROJECTS

Stable Diffusion 1.x — from random init on 2× RTX 5090 github.com/atandra2000/StableDiffusion

860M-param UNet across a 7-phase curriculum on 1.3M+ images (LAION-Aesthetic DiffusionDB/JourneyDB VGGFace2 COCO consolidation). DDPM/DDIM, Min-SNR, EMA, channels_last on Blackwell, DDP/NCCL. Best loss 0.0947 at epoch 16; epoch-42 checkpoint released on HuggingFace.

Autonomous ML Research Engineer — 15-phase multi-agent platform github.com/atandra2000/AutonomousResearcher

Paper-to-experiment end-to-end: paper analysis, repo analysis, experiment planning, code patches, training runs, statistical evaluation, autonomous looping, research reports. 23 agents, 61 tools, 186 models, 878 tests. Provider-agnostic LLM layer with self-repair.

DeepSeek-v3-Lite — faithful V3 reproduction github.com/atandra2000/DeepSeek-v3-Lite

422M params · MLA with absorption-trick inference · AuxLossFreeGate MoE · Multi-Token Prediction · MTP-as-draft speculative decoding. Companion 643-line MLA deep-dive.

FusionLLM — hybrid MLA + Gated Delta Net + MoE + MTP github.com/atandra2000/FusionLLM

415.6M active / 868.6M stored params, 24 layers, single A100 80GB. Dual-optimizer (NorMuon + CautiousAdamW), WSD + μP scheduler, 8.31B-token Chinchilla recipe.

Face Aging CycleGAN — per-layer AdaIN conditioning github.com/atandra2000/FaceAgingCycleGAN

256×256 IMDB-WIKI, 31/50 epochs on RTX 6000 Ada. Bidirectional young old with AdaIN style normalization, 3-scale PatchGAN discriminator, LSGAN + VGG-19 perceptual + L1 identity losses.

Vision-Language Model — PaliGemma-inspired, zero pretrained weights github.com/atandra2000/VisionLanguageModel

140M params · SigLIP ViT encoder + linear projector + Gemma-style GQA decoder with RoPE & GeGLU. Image patches injected at [IMG] tokens; trained end-to-end on COCO 2014 on a single P100.

EDUCATION

B.Tech, Civil Engineering 2020 – 2024

Heritage Institute of Technology, Kolkata · heritageit.edu

Civil Engineering for the math, structures, and optimization; deep learning self-taught in parallel, from raw PyTorch up to frontier LLM, diffusion, multimodal, and agentic architectures.