

A K M Asifuzzaman

AI/ML Researcher & Full-Stack AI Engineer

Chattogram, Bangladesh asif@asifzaman.dev asifzaman.dev github.com/A-K-M-Asifuzzaman
linkedin.com/in/a-k-m-asifuzzaman-6027442a7 huggingface.co/AsifZaman1912

SUMMARY

AI/ML engineer who takes systems from data audit through calibrated model to deployed interface. Domain-adapting CLIP over 95,000 scientific figures moved median retrieval rank from 83 to 26 (exact McNemar $p \approx 7 \times 10^{-31}$); a Food-101 ensemble reaches 97.16% with test ECE 0.006 after calibration; a 10-service RAG platform isolates tenants with Postgres row-level security. Evaluation is part of the build — grouped splits, held-out gates, temperature scaling, bootstrap confidence intervals. Founder of TrinolIT Solutions.

EXPERIENCE

Founder, TrinolIT Solutions

Sep 2026 - present

Dhaka, Bangladesh (remote)

Software and AI studio building bilingual (Bangla/English) business software. Delivered Nexora AI, a multi-tenant ERP/POS platform; Trinol OS and a Chromium-based browser are in development.

SELECTED PROJECTS

RoboDesign Atlas — Cross-Modal Retrieval over RobotDesign1M

2026 [Live ↗](#)

PyTorch · CLIP ViT-B/32 · FAISS · ONNX Runtime Web · FastAPI · Next.js 16 · Hugging Face Hub · uv

- Domain-adapted CLIP to scientific figures by unfreezing only the top four transformer blocks of each tower plus projections and temperature (27.5% of parameters, 58 min on an M3 Pro) under symmetric InfoNCE, lifting text→image R@1 from 12.40 to 17.24 and R@10 from 27.18 to 38.34 on the official 5,000-candidate test split — median rank 83 → 26, significant under McNemar's exact paired test at $p < 1e-12$.
- Audited the corpus and published the finding: only 5.58% of captions match robotics terminology exclusively, 50.0% match none of seven engineering lexicons, and 60,000 rows derive from just 228 instruction templates — stable to 1.3 points across all six config/split combinations, so corpus-wide rather than split skew. Re-ran the full pipeline on the 19,468-template conversations config to prove the adaptation learns the visual domain rather than caption boilerplate, reproducing the relative R@1 gain to within 0.06 points.

FoodGenome AI — Calibrated Food-101 with a Conformal Guarantee

2026 [Live ↗](#)

PyTorch · SigLIP-SO400M · EVA-02-L · DINOv2-L · ONNX Runtime · FastAPI · Next.js · Firebase

- Cached three backbones' embeddings over all 101,000 images in one 20-hour pass so every downstream ablation ran in seconds, which is what made the negative results affordable to establish rather than assume.
- Shipped a probability-average ensemble of SigLIP-SO400M and EVA-02-L at 97.16% top-1, calibrated by temperature scaling fitted on held-out data (test ECE 0.0514 → 0.0062) and wrapped in split-conformal prediction with the $(n+1)/n$ finite-sample correction for 99.56% measured coverage at 1.54 candidates per image.

DocFlow AI — Document Intelligence with Click-to-Source Citations

2026 [Live ↗](#)

FastAPI · SQLAlchemy 2 · Postgres + pgvector · ARQ · PyMuPDF · Tesseract · rapidfuzz · Next.js

- Built one provenance chain — document → page → character span → bounding box → fact → task — so every extracted deadline highlights its exact source sentence; quotes are located by fuzzy span alignment that widens from the claimed page to the whole document only as needed.
- Measured citation verification at 1.0 end to end on a clean digital PDF; because a fabricated fact cannot produce a quote that matches real page text at ≥ 95 , the rate doubles as a hallucination metric needing no judge model.

NWP-Core — A Transformer Trained from Scratch, Running in the Browser

2026 [Live ↗](#)

PyTorch · ONNX Runtime Web · WebAssembly · Next.js · Supabase · Postgres RLS

- Trained a 19.5M-parameter decoder (RMSNorm, RoPE, SwiGLU, grouped-query attention) from scratch on 262M tokens, choosing the shape by measured throughput rather than by guess.
- Reached perplexity 28.3 and top-5 accuracy 60.7% held out; shipped int8 because it cost 0.27 points of top-5 for a 4× smaller download (103.5 MB → 26.6 MB), with the ship gate reading top-5 rather than perplexity.

Beacon AI — Verifiable Meeting & Lecture Intelligence

2026 [Live ↗](#)

Next.js 15 · FastAPI · Tauri v2 · Postgres + pgvector · Apache AGE · ARQ · faster-whisper · pyannote 3.1 · BGE-M3 · Redis

- Built twelve idempotent ARQ stages keyed on stage:session_id:input_hash so any stage replays without duplicating data — 16 kHz Opus capture, whisper large-v3 + WhisperX for word-accurate offsets, pyannote 3.1 diarization, PII redaction, topic-shift chunking, and BGE-M3 dense + sparse embedding — finishing a 60-minute session in under four minutes p95 on a single A10G.
- Kept relational, vector (pgvector/HNSW), full-text and graph (Apache AGE) work inside one Postgres 16 instance instead of three databases, fusing dense ⊕ FTS ⊕ entity 1-hop graph by Reciprocal Rank Fusion and reranking with a BGE cross-encoder to eight spans, with provenance travelling alongside the text so no chunk is flattened into a prompt without its chunk_id.

NeuraHub AI — Multi-Tenant AI Workspace Platform

2026 [Live ↗](#)

Next.js 15 · React 19 · FastAPI (10 microservices + gateway) · Qdrant · Neo4j · PostgreSQL (RLS) · OpenAI · Expo · Docker · OpenTelemetry

- Architected a uv-managed Python monorepo of 10 clean/hexagonal FastAPI microservices (api → application → domain, adapters behind ports) behind a reverse-proxy API gateway that bridges Firebase web sessions into RS256 org-scoped JWTs, serving 8 product modules from one platform.
- Built GraphRAG retrieval fusing Qdrant dense vectors, BM25 keyword search, and Neo4j call-graph expansion via Reciprocal Rank Fusion — ranks, not incomparable scores — then a lexical rerank for precision, returning answers cited to file:line with evidence tagged vector / graph / fused.

Recurra — Thyroid Cancer Recurrence, Estimated Without the Leak

2026 [Live ↗](#)

Python · FastAPI · SQLAlchemy · PostgreSQL · SHAP · React · Vite · Docker

- Found by profiling that the dataset's 'Response to initial therapy' — recorded after treatment — leaks the outcome (solo AUC 0.979); trained under named feature tracks with the model registry refusing to serve the leaking one.
- Reported test PR-AUC 0.892 on features known at diagnosis against a 0.286 base rate, rather than the ~98% accuracy common for this dataset; benchmarked 120 configurations with paired significance tests over identical folds.

DermAI — Cascaded Skin Triage with Retrieval-Grounded Guidance

2026 [Repo ↗](#)

PyTorch · timm (EfficientNet-B0, MobileNetV3, ConvNeXt) · CLIP · sentence-transformers · FastAPI · React 19 · Vite 8 · gpt-5

- Proved the benchmark before trusting it: a random forest on image dimensions and file size alone — no pixels loaded — scored 99.0% on skin-vs-non-skin, and only 28,848 of 54,000 images were real photos rather than vendor rotations. Fixed with base-photo-identity grouped splits, canonicalisation (384², uniform JPEG, EXIF stripped) and a 27 → 21 class taxonomy merge, collapsing the probe to 62.5% against a 62.6% majority floor while the trained gate holds 99.6% balanced accuracy.
- Built a two-exit inference cascade a single softmax cannot express — a MobileNetV3 OOD gate that rejects non-skin and a 30%-confidence abstain threshold — over an EfficientNet-B0 classifier reporting 0.724 balanced accuracy [0.706, 0.741] and 92.2% top-3 across 21 classes at ECE 0.023 after temperature scaling, with bootstrap CIs, per-skin-tone stratification and Grad-CAM.

ALSO BUILT · ChatgaiyyaLM · Kotodama 言靈 · OmniMind · Kage · ShikkhaAI · ResearchMate AI · TaskFlow OS · DevMind AI · DevMesh AI · ImagenIQ · DataSense AI V3 · EmotiVoice AI · AutoInspect AI · PriceRadar · Meridian · CosmoMinds · Spam Detector · CineMatch AI — case studies at asifzaman.dev/projects.

EDUCATION

BSc in Computer Science and Engineering

2023 - February 2027 (expected)

East Delta University, Chattogram, Bangladesh

CGPA 3.79 / 4.00

TECHNICAL SKILLS (DEMONSTRATED IN PROJECTS)

AI / MACHINE LEARNING

Python · PyTorch · Hugging Face Transformers · PEFT / LoRA fine-tuning · scikit-learn · XGBoost / LightGBM · LSTM / RNN · ONNX export & in-browser inference

COMPUTER VISION & MULTIMODAL AI

CNNs (ResNet, DenseNet, EfficientNet) · Vision transformers (SigLIP, EVA-02) · CLIP / contrastive image-text models · Vision-language model evaluation · YOLO object detection · Grad-CAM attribution

LLMS, RAG & AI AGENTS

Retrieval-augmented generation · Hybrid dense + BM25 retrieval · Cross-encoder reranking · Multi-agent orchestration · OpenAI API · Anthropic API · Speech recognition (Whisper)

BACKEND & APIS

FastAPI · Node.js / Express / NestJS · WebSockets / Socket.IO · Task queues (Celery, ARQ) · Go

FRONTEND ENGINEERING

React · Next.js · TypeScript · Tailwind CSS · three.js / WebGL · Flutter · WebRTC

DATABASES & RETRIEVAL

PostgreSQL (incl. row-level security) · Qdrant · FAISS · Redis · Neo4j / graph retrieval · MongoDB · Pandas / NumPy

INFRASTRUCTURE & MLOPS

Docker · Hugging Face Hub & Spaces · CI with GitHub Actions · OpenTelemetry / observability · Vercel / Render deployment · Firebase

CERTIFICATIONS

Docker for Machine Learning – CampusX, Aug 2026 · FastAPI for Machine Learning – CampusX, Jun 2026 · Deep Learning: Beginner to Advanced – Codebasics, Jan 2026 · Explainable AI (XAI) – CampusX, 2025 · AI Engineer Core: LLM, RAG, QLoRA, Agents – Udemy, 2025 · NLP with Python – Udemy, 2025 · Deep Learning A-Z 2025 – Udemy, 2025 · A Deep Understanding of Deep Learning – Udemy, 2025 · Machine Learning A-Z (AI, Python & R) – Udemy, Apr 2025 · Supervised ML: Regression & Classification – DeepLearning.AI / Coursera, Feb 2025 · Python for Data Science, AI & Development – IBM / Coursera, Feb 2025 · Master ML for Data Science – Codebasics, Jan 2025 · Complete Web Development – Programming Hero, Jun 2024