

MUHAMMAD ASJAD

Islamabad, Pakistan · m.asjad2401@gmail.com

asjad.apostrion.site · github.com/asjad2401 · arxiv.org/a/asjad_m · linkedin.com/in/muhammad-asjad-b992542a6

RESEARCH INTERESTS

Empirical evaluation of machine-learning systems: whether standard benchmarks and metrics measure the construct they claim to. Related interests in program analysis and binary security, vision-language evaluation, and reproducible research methodology.

EDUCATION

BSc in Computer Science — National University of Sciences and Technology (NUST)

Aug 2023 – 2027 (expected)

School of Electrical Engineering and Computer Science (SEECs), Islamabad, Pakistan.

Coursework: Deep Learning · Machine Learning · Artificial Intelligence · Cyber Security · Information Security · Computer Networks · Compiler Construction · Operating Systems · Parallel & Distributed Computing · Design and Analysis of Algorithms · Data Structures & Algorithms · Theory of Automata & Formal Languages · Computer Architecture · Linear Algebra · Probability & Statistics · Discrete Mathematics.

PUBLICATIONS AND PREPRINTS

A Memorization Floor for LLM Refinement of Decompiled Code

arXiv:2609.17236 [cs.SE]

M. Asjad (single author), 2026. Under review at Empirical Software Engineering (Springer).

Pre-registered study — analysis plan committed to version control before any data was collected — of what a large language model contributes when refining Ghidra decompiler output. On functions written after registration, the model recovers identifiers no better from intact input than from input whose identifiers have been scrambled: it reproduces a naming prior rather than reading the code. Replicated against a second vendor’s model on byte-identical inputs; all twelve contrasts across two models and two corpus tiers span zero. Reports five transferable instrument findings, including a reassembly harness whose bugs manufactured a 20-point correctness deficit by penalising the model for obeying its own prompt, and an LLM readability judge that ranks arms as a blind human rater does while reporting a gap 2.8x larger.

Replication package — corpus, pipeline, pre-registered analysis plan, dated research log, raw API responses: doi.org/10.5281/zenodo.21968878

What Does Emotion Accuracy Measure? A Pre-Registered Study of Lexical Provenance in Emotion-Conditioned Image Captioning

Manuscript complete

M. Asjad (single author), 2026.

Pre-registered study of what emotion-conditioning accuracy measures when the training captions are generated by a language model. One fixed ClipCap-style architecture trained across eight caption corpora differing only in provenance, structure and volume (48 runs); all arms scored by a single frozen DistilRoBERTa classifier, with every accuracy reported beside a top-25-keyword baseline recomputed on that arm’s own evaluation set. Raw accuracy tracks lexical provenance rather than emotional register: the most keyword-stereotyped corpus leads by 10.6 accuracy points and trails by 2.5 on the margin, both directions Holm-significant. A training-free baseline emitting five fixed strings and never seeing the photograph scores 0.80 — above five of eight trained arms — at a margin of -0.20 . Validated by three blind human raters, an off-distribution check on 148 hand-written captions, and an artifact ablation.

RESEARCH EXPERIENCE

Final-Year Thesis — IRES: Intelligent Reverse Engineering System

Aug 2026 – present

Supervisor: Dr. Khurram Shehzad, SEECs NUST. LLM-assisted binary decompilation. The pre-registered study above is the project’s first output.

Independent Research — Generalisation to Unseen Attack Categories in Network Intrusion Detection

2026

Main study — NSL-KDD, canonical train/test split. Eight tuned model families x four held-out attack categories x five seeds, with an explicit negative control and a positive control establishing each category’s achievable ceiling, so the reported quantity is novelty cost (ceiling minus unseen recall) rather than raw recall. Withholding an entire attack family costs 0.19–0.23 recall regardless of architecture (tree vs. neural: 0.207 vs. 0.214, $p = 0.58$); unseen families retain 50–79% of achievable recall. Robustness correlates with same-task skill at $\rho = +0.738$ ($p = 0.037$) and with different-task skill at $\rho = -0.714$ ($p = 0.047$) — same models, same run, opposite publishable conclusions.

Extension — open-set detection benchmark. Twelve scorers, including kNN-distance, energy, ODIN, OpenMax and Deep SVDD, across 23 held-out families on NSL-KDD, UNSW-NB15 and CICIDS2017, tuned and untuned. No purpose-built open-set method beats a plain classifier’s own confidence on either modern benchmark (0 of 12, and 0 of 6 after a 30-setting sweep per dataset), while 7 of 12 beat it on NSL-KDD; the two modern benchmarks rank the methods at $\rho = +0.825$ ($p < 0.001$) and NSL-KDD correlates with neither. The mechanism is attack–benign separation, computable from unlabelled geometry before any detector is trained: fitted on two benchmarks it removes 32% of the error in forecasting the third, with the skill concentrated in distance-like scorers (+0.497) rather than confidence-based ones (+0.053). Predictions for CICIDS2017 were committed to version control before any CICIDS detector was trained; the rank claim replicated out of sample and the numeric-threshold claim was falsified 1 of 2, reported as a falsification with no threshold adjustment or re-scoring. Three competing explanations were tested; two rejected. Results complete across both studies; manuscript in preparation.

Upstream Defect Reports — codealign (Dramko et al., TOSEM 2025)

2026

Identified and reported three reproducible defects in a published graded equivalence-checking tool at HEAD, after an independent study found it did not discriminate on optimised decompiler output (informative-pair rate 0.092 against a pre-registered floor of 0.20). Two of the three reduce to a few lines of plain C with no decompiler involved. github.com/asjad2401/codealign-repro

Machine Learning Research Intern — TUKL-NUST R&D Center

May – Sep 2024

Slum mapping via semantic segmentation of very-high-resolution satellite imagery sourced from SUPARCO, Pakistan’s national space agency. Reproduced and benchmarked leading U-Net-based approaches against the project dataset; findings contributed to an internal research publication.

TECHNICAL SKILLS

Languages — Python, Rust, C, C++, Java, SQL, Dart, JavaScript/TypeScript

ML and research — PyTorch, TensorFlow, scikit-learn, Hugging Face Transformers, XGBoost/LightGBM, ONNX Runtime, TensorRT, CLIP, LoRA/PEFT; experimental design, pre-registration, cluster bootstrap, permutation tests, Holm–Bonferroni correction, power analysis, inter-rater reliability (Krippendorff’s alpha, Cohen’s kappa)

Systems and tooling — Linux, Git, Ghidra, GDB, Docker, FastAPI, MongoDB, LaTeX, Cloudflare; distributed systems, compiler construction, OS internals

OPEN-SOURCE CONTRIBUTIONS

Merged pull request — PyO3/rust-numpy #550

May 2026

Fixed a correctness bug in PyArray’s exact-type checking that permitted invalid casts and failed to reject array subclasses; implemented a configurable fix per maintainer review, with tests and a changelog entry. Merged into the actively maintained crate. github.com/PyO3/rust-numpy/pull/550

rustnum — NumPy-compatible numerical library in Rust

2026 – present

Drop-in API for ML activation functions (sigmoid, softmax, ReLU) backed by Rust; 1.9–3.3x faster than NumPy through multi-core parallelisation. github.com/asjad2401/rustnum

burn-python — Python inference frontend for the Burn deep-learning framework

2026 – present

Loads ONNX models and runs inference with pure NumPy in and out, requiring no Rust knowledge from the user. CI configured. github.com/asjad2401/burn-python

patina — a Unix shell written from scratch in Rust

2026 – present

Own REPL, tokenizer and process management via raw fork/execv/waitpid, building toward reproducible, replayable terminal sessions with SHA-256-hashed command provenance. github.com/asjad2401/patina

PROFESSIONAL EXPERIENCE

AI Intern — Sketric Solutions

Jul – Sep 2026

Three-person team operating a human-in-the-loop annotation pipeline for a real-time hazard-detection computer vision system; reviewed and corrected model annotations across roughly 5,000 images to iteratively retrain and improve accuracy.

AI Engineer (part-time) — Veron Solutions

Jun – Sep 2024

Five-person team building a YOLO and DeepSORT pipeline for continuous vehicle detection and tracking in drone footage; curated and labelled public aerial-imagery datasets for training.

Founder — Apostrion

Jul 2026 – present

Founded an AI and software consultancy and built its first product end-to-end: an AI customer-support chatbot trained on a business’s own FAQs, deployable as a hosted link, website widget or WhatsApp bot. Covers backend API, chat product and marketing site.

SELECTED PROJECTS

- OFFLOAD-HPC (Java) — distributed compute grid offloading matrix multiplication, hash cracking and image processing from mobile devices to a LAN of PCs. UDP-based node discovery and Bully-algorithm leader election give self-organising grid membership that survives node churn.
- VCSDSA (C++) — a local version-control system implemented from scratch: custom object store, staging area, branches and commits.
- LocalQwenModel (Python) — fully offline coding agent using Ollama (qwen3:8b) with ChromaDB retrieval-augmented generation and a FastAPI + SSE streaming interface.
- PKR-Detection (Python) — real-time YOLOv11n detector for Pakistani banknotes, built for assistive devices for visually impaired users.
- Voice-Assisted Calling (Dart/Flutter) — fully voice-controlled dialer for blind users: speech-to-text dialing, fuzzy contact matching, text-to-speech caller announcement, Android default-dialer integration.
- OptimalAI (Python/TypeScript, 4-person team) — owned backend and LLM integration; FastAPI/MongoDB service with a scraping pipeline and Gemini via LangChain, React 19 frontend, Playwright end-to-end tests.
- Agri_intel (Python) — district-wise wheat-yield modelling for Punjab combining government statistics, MODIS NDVI satellite data and NASA POWER climate data with gradient boosting; Streamlit dashboard.
- Comparative NLP study — article classification via TF-IDF with softmax regression versus BERT embeddings; and a one-hidden-layer neural network implemented from scratch in NumPy.

Languages: English (professional working) · Urdu (native) · German (elementary)