

MAINUL HOSSAIN

yaadmainul@gmail.com | Fairfax, VA

Website: yaad96.github.io | LinkedIn: [mhyaad](#) | GitHub: [yaad96](#) | Google Scholar: [Mainul Hossain](#)

PROFILE

CS PhD student researching how to make AI coding agents reliable and trustworthy: giving LLM agents the project context they need to repair code, and helping developers oversee what agents produce. Designs agents, benchmarks them at scale, and runs user studies. 8 peer-reviewed papers (58 citations, h-index 4), including a first-author ASE 2026 paper. Seeking a Summer 2027 software engineering or ML engineering internship.

EDUCATION

Ph.D., Computer Science | George Mason University, Fairfax, VA | GPA: 3.95/4.00 | *Expected May 2028*

M.Sc., Computer Science | George Mason University, Fairfax, VA | GPA: 3.95/4.00 | *May 2026*

Distinguished Academic Achievement Award

B.Sc., Computer Science and Engineering | BUET, Dhaka | GPA: 3.35/4.00 | *February 2021*

SKILLS

Languages: Python, Java, TypeScript/JavaScript, C++, SQL

LLMs: RAG, Prompt & Context Engineering, LLM Agents, Tool Calling, Guardrails, LLM Evaluation, SFT, DPO

ML/AI: PyTorch, TensorFlow, Hugging Face Transformers, LoRA/PEFT, Scikit-learn, NLP, Computer Vision

Web/Backend: React, Node.js, Django, FastAPI, Firebase, PostgreSQL, Redis, Celery

Tools: Docker, Git, AWS, OpenAI API, REST APIs, StrykerJS, Vitest

Data Analysis: Pandas, NumPy, Matplotlib, Jupyter, Statistical Analysis (paired t-test, Wilcoxon, McNemar)

Research: Mixed-Methods Study Design, Qualitative Coding, Thematic Analysis, Mutation Testing, User Studies

EXPERIENCE

Software Engineering Intern | *Innovation Lab, MetroStar Inc., Fairfax, VA, Summer 2026*

- Built a knowledge-base-first RAG pipeline for Sapphire, MetroStar's in-house AI assistant: it grounds each answer in company knowledge, then applies the LLM's reasoning and general knowledge as needed.
- Engineered token-efficient prompts and context across GPT and self-hosted open-weight Llama models, with guardrails that make Sapphire refuse off-topic requests and queries exposing sensitive company data.

Software Engineering Intern | *Innovation Lab, MetroStar Inc., Fairfax, VA, Summer 2025*

- Built Falcon end to end, a full-stack MVP for drone mission planning and monitoring: drone selection, mission templates, waypoint planning, and live mission execution.
- Developed reusable form components (takeoff, landing, waypoint, drone configuration), map-based waypoint rendering, and the database and backend that propagate mission data in real time.

Graduate Teaching Assistant | *George Mason University, Fairfax, VA, Aug 2023 - Present*

- Deliver lectures, hold office hours, grade assignments, and lead lab sessions for Object-Oriented Programming (CS 211), Introduction to Low-Level Programming (CS 262), and Database Systems (CS 550).

Software Developer | *BANBEIS, Dhaka, Bangladesh, Sep 2022 - Jul 2023*

- Developed a web-based public examination processing system for Bangladesh using computer vision (OMR processing) to detect handwritten digits.
- Resolved data anomalies ensuring accurate result processing and migrated examination records to a centralized server.

Research Assistant | *BUET, Dhaka, Bangladesh, Feb 2019 - Sep 2022*

- Co-authored peer-reviewed papers in HCI, AI/ML, and data science (IEEE Access, Heliyon, Scientific Reports, IEEE ICDM, ACM ICTD) on survey data integrity, vehicle detection, and rail-track monitoring.

RESEARCH AND PROJECTS

DesignFix: Context-Augmented Repair of Design Rule Violations

Published at ASE 2026 (first author) / 2024 - 2025 / 2024 Summer PhD Research Award

- Built DesignFix, an IDE-integrated, human-in-the-loop LLM tool that repairs design rule violations, grounding each fix in the rule description, a compliant example from the same project, and surrounding code.

- Designed context-augmented, two-step prompting: up to 90% repair success (+26 points over baseline), 89% explanation correctness, and 90% unrelated-code preservation in 9,450 GPT-4o trials.
- Ran a within-subjects study (n=14) vs a browser-LLM control: 36% less task time (10.0 vs 15.7 min, $p < 0.001$), correct repairs up from 43% to 93%, and significantly lower workload on 5 of 6 NASA-TLX dimensions.
- Showed that DesignFix's primary value is codebase navigation, not code generation (first code edit at 2.3 vs 5.8 min), and that the Cline coding agent repaired only 3 of 6 task types with 12.1× the tokens.

AgentFlake: An Agentic Framework for Repairing Flaky Tests

2026 - Present / Full paper under review

- Co-led development of AgentFlake, the first agentic framework for repairing flaky tests: its LLM agent gathers context via read-only tool calls and decides when to patch, and each fix must pass 100 reruns.
- Benchmarked AgentFlake on 164 flaky tests from 53 Java projects: 158 repaired (96.3%) vs FlakyGuard 74.4%, FlakyDoctor 50.0%, Claude Code 90.2%, Codex 79.9%, at 30% lower cost and 72% less time than FlakyGuard.

Developer Metacognition, Trust Calibration, and Oversight in LLM-Assisted Software Testing

2026 - Present / Ongoing

- Designed and led a convergent mixed-methods study (n=19) investigating how developers interact with LLM coding assistants (Cursor, Claude Code, Amazon Q), analyzing 26,160 words of structured reflections, 19 codebases, and 3,269 test cases across 175 test files.
- Built an automated mutation testing pipeline (StrykerJS + Vitest) across all participant codebases, revealing that developers reporting 85-100% branch coverage achieved only 44-82% mutation scores, quantifying a trust calibration gap of up to 56 points.
- Developed a taxonomy of six developer oversight strategies and traced how strategies shift across task complexity, finding that multi-method oversight predicts calibration accuracy while single-metric reliance systematically produces overconfidence.
- Conducted reflexive thematic analysis (Braun & Clarke, 2022) producing seven themes on human-AI collaboration, including how LLM-generated tests substitute structural completeness for logical depth and how coverage metrics function as psychological closure signals.
- Performed automated test smell detection (2,613 instances, 16 smell types), finding that smell type predicts mutation effectiveness more strongly than smell quantity.

Quantitative Currency Evaluation for Visually Impaired Users

[*RDM Workshop @ IEEE ICDM 2025*](#)

- Built a three-stream vision pipeline for Bangladeshi banknotes (denomination, usability, counterfeit detection); a 287K-parameter CNN on an 82,512-image dataset reached 0.9342 F1 and 0.99 counterfeit-detection accuracy.

STRIP: Supervised Training with RL for Irrelevant Prompt Removal in Math Reasoning

- Proposed STRIP, a two-phase framework that trains an LLM to remove irrelevant sentences from noisy math word problems, producing a cleaned prompt in one inference call.
- Implemented STRIP on Llama 2 7B using LoRA adapters (r=64, alpha=128) with 8-bit quantization, trained on GSM-IC with supervised fine-tuning (SFT) followed by Direct Preference Optimization (DPO).
- Designed a multi-component reward function with asymmetric penalties, achieving 0.9927 filtering F1 with SFT and 0.9894 with DPO on GSM-IC.

Bulk Invoice Download and Export

[*https://marketplace.stripe.com/apps/bulk-actions*](https://marketplace.stripe.com/apps/bulk-actions)

- Published a Stripe Marketplace app for bulk PDF invoice export; built with React, Django, PostgreSQL, Celery, Redis, Docker Compose; deployed on AWS.

Analyzing Negative Experiences in Hajj Pilgrimage

[*IEEE Access 2021 & Heliyon 2023*](#)

- Surveyed 988 pilgrims; quantified data-collector integrity; linked 21 issue types to recommendations.