Mobasshir Bhuiya Shagor

Research summary, September 2026

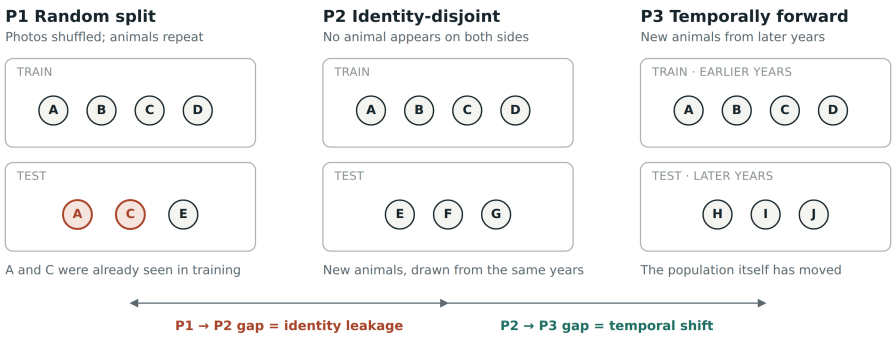
Staff ML Engineer, Graaho Technologies, Dhaka, Bangladesh · ORCID 0000-0003-3912-7650
mobasshir.bhuiya@gmail.com · mobasshirbhuiya.com · Google Scholar · GitHub

The question

I build instruments (datasets, evaluation protocols and audits) that test whether a learned system deserves the trust people give it. Personalization is where I study them, because that is where models most often act for someone rather than answer them. The thread through the work is one question: does a reported number survive the conditions it will actually meet? For me it began in 2019, with a model that estimates a cow's weight and breed from photographs, a runner-up at Bangladesh's national ICT awards. The award never asked whether it would still work a year later, on animals it had never seen, so I spent the next six years collecting the data to find out.

BoviShift: separating what a benchmark measures from what it remembers

First author; manuscript available on request. With N. H. Tonmoy (Indiana University) and Dr. M. S. U. Miah (AIUB). I collected images of 2,657 bulls sold through one livestock marketplace between 2021 and 2026, with every image linked to the animal it shows. A random split lets the same animal appear in training and test; BoviShift adds an identity-disjoint protocol and a temporally forward one, so that an accuracy drop decomposes into identity leakage and temporal shift instead of being reported as one number. The 2024 cohort is held out of the primary temporal arm because its missingness is non-random. Most personalization data is repeated observation of the same people, so the decomposition transfers directly to recommender benchmarks.



Auditing LLM-based recommenders

Sole author, in preparation.
An LLM asked why it recommended an item will give a reason whether or not that reason drove the ranking. I am building an audit that measures explanation faithfulness, hallucinated item attributes, and whether either failure falls unevenly across user groups. It runs without GPU compute, so a team can repeat it on its own system.

Where the questions come from

Nearly eight years of production ML. On AlgoRec, a metered recommendation platform, no model went live until it beat a popularity baseline on AUC. On Koronik, an agent platform serving 5,000+ daily users at Islami Bank Bangladesh, the agent refuses and hands off to a person below a confidence threshold. On Margin, an adaptive writing tutor, an LLM grades while deterministic code bounds every score it can show. Each rule is a small evaluation instrument built into a product. My research asks what it takes to make them general.

What I want to work on in a PhD

- Longitudinal benchmarks for recommender systems that separate leakage from real generalisation.
- Faithfulness and fairness audits for LLM recommenders that are cheap enough to become routine.
- Evaluation for agents that act on a user's behalf, where an unverified decision has a cost.

Earlier publications

M. B. Shagor, M. Z. H. Alvi, K. T. Hasan. CID: Cow Images Dataset for Regression and Classification. *ICCA '22*, doi:10.1145/3542954.3543018. **M. B. Shagor** et al. CAM – Categorized Affect Map. *ICCA 2020*, doi:10.1145/3377049.3377074. Full CV at mobasshirbhuiya.com.