# Intesar Tahmid

*Research Statement*

intesar3006@gmail.com · intesar-tahmid.tech · Google Scholar · October 2026

## Overview

I want to build language and vision-language models that can be trusted where a wrong answer has consequences, and that keep getting better after they are deployed. I arrived at this question through experiments, not theory. Almost every project I have worked on began as an attempt to make a model useful in Bangla or in medicine, and almost every one ended by showing me how far a model's benchmark score can sit from its real reliability. My research now pursues two connected goals: measuring what models actually catch in high-stakes settings, and using those measurements to let models improve themselves at test time, with rewards that can be verified rather than guessed.

## Measuring what models actually catch

BanglaMedVQA [1, 2] began with a plain gap. Models were already being promoted for clinical use, yet there was no medical visual question answering benchmark for Bangla, one of the most widely spoken languages in the world. We built 7,000 image–question pairs spanning five question types, from imaging modality and organ to abnormality, condition, and anatomical position. Two medical specialists independently reviewed a 1,000-sample subset, with 97% of items accepted. We then evaluated nine open, closed, and medical LVLMs under zero-shot, chain-of-thought, and LoRA fine-tuning.

What stayed with me was not the average score but its shape. On Bangla questions, Gemini 2.5 identified the imaging modality 87.5% of the time, yet answered fewer than one in four condition and position questions correctly, the questions closest to an actual diagnosis. LoRA fine-tuning lifted compact open models such as Qwen2.5-VL and MedGemma past every closed model overall (58.25% against 35.75% for the strongest closed model), but the specialized categories stayed low. Fine-tuning sharpened what models could already do; it did not create diagnostic reasoning.

That pattern shaped RxFact [3], my first-author benchmark, now under review. Rather than asking whether a model can answer a question about a medical image, RxFact asks whether it notices that a prescription image contains a clinically harmful error. Detection is a stricter test than answering. A model that reads a prescription fluently but misses a dangerous error is worse than unhelpful, because it lends confidence to a mistake.

## Adapting models cheaply, and stably

Evaluation shows where models fail; adaptation is how we respond, and in low-resource settings it has to be cheap. In LP-FT-LoRA [4], we addressed a quiet source of instability: when LoRA adapters and a new classifier head are trained together from a cold start, the head chases features that are still moving. We separated the two, first aligning the head on frozen features with a linear probe, then training only the adapters, then briefly refining both. Across five Bangla benchmarks and four compact models, the method consistently outperformed standard LoRA fine-tuning and generalized better out of distribution. An earlier shared-task system built on linear probing [5] placed in the top 15 in all three BLP-2025 hate speech subtasks. Together, these projects taught me to treat adaptation as a sequence of decisions about what to train and when, rather than a single knob.

## Defending learning against adversaries

In the Green Networking Research Group, advised by Prof. Md Abdur Razzaque, I co-led AD-SFL [6], now under review. It defends SplitFed learning by spotting poisoned clients as anomalies in activation space, then letting a reinforcement-learning policy decide how aggressively to exclude

them, without assuming how many attackers are present. The lesson carries directly into my current work: whenever a model learns from data it does not fully control, the learning signal itself becomes something to attack.

Alongside research, I work as an AI engineer at Penta Global Ltd, building agentic LLM pipelines, retrieval systems, and fine-tuned models that run in production. Engineering keeps me close to a practical fact: a deployed model meets inputs that no benchmark anticipated, so neither evaluation nor improvement can stop at release.

### Current work: learning at test time

That fact leads to the question I am working on now. Test-time reinforcement learning lets a model update itself on the unlabeled inputs it meets after deployment, guided by rewards it can check for itself. In StreamTTRL, which I lead and am preparing for ACL 2027, I study what happens when this learning is continual: when the stream of inputs shifts across languages, domains, and modalities, does a model accumulate skill, or does it forget, drift, and reinforce its own mistakes?

### Research agenda

In a PhD, I want to pursue three directions that grow out of this work.

1. **From failures to rewards.** Benchmarks such as RxFact surface specific, checkable failures. I want to turn those failures into verifiable reward signals, so that a model is trained, or trains itself, against the errors that matter most rather than against generic correctness.
2. **Continual test-time learning at scale.** I want to understand when test-time reinforcement learning accumulates skill over a changing stream and when it collapses, and to develop methods, such as calibrated rewards and selective updates, that let models keep learning across languages and modalities without forgetting what they already know.
3. **Robust learning after deployment.** A model that learns from its own test-time inputs hands an adversary a new lever: poisoned inputs can corrupt the reward itself. Building on AD-SFL, I want to study how to detect and resist such poisoning when the learner, and not only its training set, is exposed.

Across this work I have come to see evaluation and learning as two halves of one problem. We cannot improve what we cannot measure honestly, and measurement matters most when it changes how models learn. I am looking for a PhD environment where careful benchmarks and new learning methods are built side by side, toward models that deserve the trust people already place in them.

---

### Selected publications


1. R. Ahmed*, **I. Tahmid***, M. S. Hossain*, et al. Evaluating Large Vision Language Models on Bangla Medical Visual Question Answering. *Findings of ACL*, 2026.
2. R. Ahmed*, **I. Tahmid***, et al. How Good LLMs Are at Answering Bangla Medical Visual Questions? Dataset and Benchmarking. *2nd AAAI Bridge Program on AI for Medicine and Healthcare*, 2026.
3. **I. Tahmid**, A. M. R. Mazumder, M. S. Hossain, et al. RxFact: Can Large Language Models Detect Clinically Harmful Errors in Prescription Images? *Under review, ACL Rolling Review*, Oct 2026.
4. T. H. Tomal*, A. B. Uddin*, **I. Tahmid**, et al. LP-FT-LoRA: A Three-Stage PEFT Framework for Efficient Domain Adaptation in Bangla NLP Tasks. *BLP Workshop, IJCNLP-AACL*, 2025.
5. **I. Tahmid***, R. Ahmed*, et al. PentaML at BLP-2025 Task 1: Linear Probing of Pre-trained Transformer-based Models for Bangla Hate Speech Detection. *BLP Workshop, IJCNLP-AACL*, 2025.
6. T. H. Tomal*, **I. Tahmid***, P. Roy, M. Hasan, M. A. Razzaque. AD-SFL: Activation-Space Anomaly Detection and Reinforcement Learning-Based Defense Against Data Poisoning in SplitFed Learning. *Under review, Computer Networks (Elsevier)*.


* Equal contribution.