Taleef Tamsal

+1 (260) 515-9571 | taleeftamsal@hotmail.com | taleef7.github.io
linkedin.com/in/ttamsal | github.com/Taleef7 | Google Scholar | ORCID 0009-0002-7025-4842

Research Interests

Adversarial robustness and safety evaluation of large language models, with particular attention to how efficiency interventions alter safety properties, how tool-using agents evade text classifiers, and whether reported robustness reflects the model or the measurement pipeline used to score it.

Education

Purdue University Fort Wayne — Fort Wayne, IN
M.S. in Computer Science, GPA 4.00/4.00 — *Aug. 2024 – May 2026*

- Advisor: Dr. Jonathan Rusert
- Thesis: *Do Efficient Adaptations Reduce Safety? Jailbreak Robustness of PEFT vs. Full Fine-Tuning on Consumer-Accessible LLMs.* Nearly 600 attack-success-rate measurements across 25 model configurations, calibrated against a 750-sample human annotation study.

Lahore University of Management Sciences (LUMS) — Lahore, Pakistan
B.S. in Computer Science — *Aug. 2020 – May 2024*

Publications

All entries are first-authored.

Peer-Reviewed

- **Taleef Tamsal** and Jonathan Rusert. Your Judge Is a Confound: Evaluator and Attack Choice Distort Jailbreak-Safety Measurement for Fine-Tuned LLMs. *AdvML-Frontiers × CoTMA Workshop*, COLM 2026. Accepted.
- **Taleef Tamsal** and Jonathan Rusert. PFW Task 8 at SemEval-2026 Task 8: Lightweight Tri-Fusion Retrieval with Prompt-Engineered Faithful Generation for Multi-Turn RAG. *Proceedings of the 20th International Workshop on Semantic Evaluation*, pages 1526–1532, ACL, 2026. Code.
- **Taleef Tamsal** and Jonathan Rusert. PFW at SemEval-2026 Task 6: Multi-Seed DeBERTa Ensembles for Political Response Clarity and Evasion Classification. *Proceedings of the 20th International Workshop on Semantic Evaluation*, pages 1518–1525, ACL, 2026. Code.

Under Review

- **Taleef Tamsal** and Jonathan Rusert. Adaptation, Not Algorithm: PEFT and Full Fine-Tuning Degrade Jailbreak Safety Equivalently Across Five Open-Weight LLMs. ACL Rolling Review, May 2026 cycle, committed to **EMNLP 2026**. Decision pending. Code.

In Preparation

- **Taleef Tamsal** and Jonathan Rusert. Shapley Attribution over Tool Availability for Agentic Adversarial Attacks on Toxicity Classifiers. Targeting ACL Rolling Review, October 2026 cycle.

Research Experience

Graduate Researcher, AI and LLM Safety — Fort Wayne, IN
Purdue University Fort Wayne, advised by Dr. Jonathan Rusert — *Sep. 2025 – May 2026*

- Evaluated how LoRA, QLoRA, 4-bit quantization, and full fine-tuning affect jailbreak robustness across five open-weight families (Llama-3.1, Qwen-2.5, Qwen3, Phi-4, Gemma-2) under PAIR, DeepInception, ArtPrompt, AutoDAN, and ActorAttack.
- Established by two one-sided tests that LoRA and full fine-tuning degrade black-box safety equivalently within ±5pp under fixed-seed retraining, while QLoRA separates as a higher-variance method.
- Ran a 750-sample blind annotation study calibrating three automated evaluators, showing that a widely used LLM judge runs at a 45.7% false-positive rate against human consensus versus 8.6% for a fine-tuned classifier, which inflates measured ΔASR roughly threefold and reorders method rankings.
- Showed that black-box and white-box attacks return directionally opposite verdicts for the same method, and that a published defense's headline +58pp protection collapses to +0.5pp against a matched-seed baseline.
- Released an open attack and defense harness for reproducible robustness evaluation on consumer-sized models.

Research Collaborator, Query-Efficient Adversarial NLP Fort Wayne, IN
Purdue University Fort Wayne, Dr. Jonathan Rusert's group *June 2025 – Aug. 2025*

- Mapped the selection and replacement stages of black-box text attacks (TextFooler, TextBugger, and related methods) onto a proposed query-efficient attack to identify which components admitted optimization.
- Built comparison tables across datasets, victim models, query budgets, and attack success rates for BinarySelect-style query-efficient attacks.

Ongoing Research

Agentic Adversarial Attacks with Tool-Level Attribution *2026 – Present*

- Building a ReAct-style attacker over a discrete tool registry that evades toxicity classifiers under a hard query budget, evaluated against four self-trained victims per dataset above 0.85 macro-F1 (ModernBERT-large, RoBERTa-large, DeBERTa-v3-large, QLoRA fine-tuned Gemma-3-12B) on HateXplain and Civil Comments.
- Computing exact Shapley values over all 2^7 tool coalitions for two independent planners, with bootstrap confidence intervals and interaction indices, to attribute attack success to individual operators.
- Measured instrument failure on untouched text, a 23% retention-classifier failure rate on unedited originals and a 0.20 to 0.97 spread across four toxicity instruments, indicating that reported validity depends heavily on the measuring apparatus.
- Running under a git-tagged preregistration with a full deviation log, machine invariant auditing, and mandatory human reading of sampled adversarial text per experimental cell.

Shared Task Participation

UNLP 2026, Ukrainian Multi-Domain Document Question Answering *2026*

- Built a retrieval and answering pipeline combining BGE-M3 dense retrieval, a Qwen3-0.6B cross-encoder page reranker, and a quantized MamayLM-Gemma-3-12B answerer under a 9-hour offline compute limit.
- Formalized risk signals for predicting generalization failure after every offline improvement regressed on the hidden leaderboard.

UNLP 2025, Social-Media Manipulation Detection *2025*

- Built a joint system for multi-label classification of 10 manipulation techniques and span-level detection over 9.5K Ukrainian Telegram posts, using rationale-guided training and class-balanced focal loss to make predictions explainable.

SemEval-2025 Task 10, Multilingual Narratives *2024 – 2025*

- Built XLM-R and BERT pipelines for entity-role framing, narrative classification, and narrative extraction across five languages, reaching roughly 8× improvement over the shared baseline on internal validation.

Industry Experience

Medical Informatics Engineering Fort Wayne, IN
Software Developer *May 2026 – Present*

- Sole developer on production services behind real-time REST APIs, backed by 785 automated tests and validated against official certification test cases.
- Designed a resumable batch pipeline over 70,000 PostgreSQL records with bounded memory and per-record failure isolation.

Medical Informatics Engineering Fort Wayne, IN
Development Intern, NLP and Retrieval-Augmented Generation *May 2025 – Aug. 2025*

- Built a containerized retrieval-augmented search platform giving citation-backed semantic search over long-form enterprise documents, using Node.js ingestion, hybrid OpenSearch (BM25 and vectors), and an MCP interface.
- Added RAGAS and TruLens as pre-deployment quality gates on grounding, faithfulness, relevance, and latency, measuring 85%+ relevance across model and retriever configurations.
- Indexed 3,000+ issue records and prototyped a clinical-notes workflow so engineers and clinicians could query prior fixes in natural language.

Teaching Experience

Teaching Assistant, Algorithms (CS 310) LUMS
Led tutorials and office hours, graded assignments and quizzes. *Aug. – Dec. 2023*

Teaching Assistant, Introduction to Blockchain (CS 3812) LUMS
Updated course assignments, ran weekly tutorials and office hours, assisted with grading. *Jan. – May 2023*

Teaching Assistant, Topics in Blockchain Technology (CS 589) LUMS
Facilitated weekly graduate paper discussions and supervised semester projects. *Aug. – Dec. 2022*

Teaching Assistant, Computational Problem Solving (CS 100) LUMS
Ran tutorials and office hours for first-year students, graded labs and quizzes. *Aug. – Dec. 2022*

Mentoring and Service

Mentor, LUMS Alumni Mentorship Program *July 2024 – Present*

- Advise undergraduates on graduate applications (statement of purpose, program targeting, timelines), the transition into CS research, and research career pathways in the US.

Event Lead, Pakistani Student Association, Purdue University Fort Wayne *Jan. 2025 – May 2026*

- Planned and ran cultural and social events for the association, coordinating venues, budgets, volunteers, and promotion across the semester calendar.
- Built events that welcomed students from outside the community as well as within it, giving incoming international students an early point of contact on campus.

Mentor, GradApp Lab *July 2024 – Jan. 2025*

- Mentored undergraduates and early-career CS students from Pakistan through the graduate application process and the move into research.

Peer Ambassador for Social Support, LUMS *Sep. 2023 – May 2024*

- Mentored first-year students on academic, social, and personal well-being.

Technical Skills

Languages: Python, C/C++, SQL, TypeScript, Bash, R, Go, Scala

ML and NLP: PyTorch, Hugging Face Transformers, PEFT (LoRA, QLoRA), bitsandbytes, vLLM, DeepSpeed, TensorFlow, scikit-learn, spaCy, LangChain

Evaluation and statistics: equivalence testing (TOST), bootstrap confidence intervals, nonparametric hypothesis testing, multiple-comparison correction, mixed-effects models, inter-annotator agreement, calibration, human annotation design, preregistration

Retrieval and RAG: OpenSearch, FAISS, BM25, SPLADE, BGE-M3, hybrid retrieval, cross-encoder reranking, RAGAS, TruLens

Infrastructure: SLURM and HPC clusters, Docker, Git, GitHub Actions, PostgreSQL, Weights & Biases, AWS, Azure, Google Cloud Platform, Linux