ERFAN BAYAT, MS

+39 344 508 4954 ⋄ Turin, Italy ⋄ erfanbyt13@gmail.com
Linkedin ⋄ Github

PROFESSIONAL SUMMARY

AI Research Engineer with published work in LLM safety and multimodal systems. Expertise in adversarial robustness, NLP, and speech processing pipelines. Experience collaborating with Amazon on speech LLM research.

EDUCATION

Master of Science in Data Science and Engineering, Politecnico di Torino 2023–2025
Thesis: Adversarial RAG-based approach to counter-narrative generation
Board Member, MALTO (Machine Learning at Polito)

Bachelor of Computer Science, Islamic Azad University, Science and Research Branch 2019 - 2023

PUBLICATIONS

Led research and co-authored "**Dual Teachers for Unlearning Sensitive Content in LLMs**" published in the SemEval 2024 workshop proceedings (ACL Anthology, 2025). Developed a novel Dual-Teacher framework leveraging Competent and Incompetent Teachers with KL divergence minimization to remove unwanted information from LLMs while preserving model utility and adversarial robustness. here

EXPERIENCE

Research Scholar - Speech LLM Optimization July 2025 - Present
Politecnico di Torino (Amazon-Sponsored Scholarship) Turin, Italy
Selected for competitive Amazon-sponsored research to advance Speech LLM technology in collaboration with Amazon AGI. Developed a scalable pipeline leveraging LLM, ASR, and TTS components to generate 50K+ synthetic conversational datasets for dialogue system training. Implementing reinforcement learning from human feedback (RLHF) methods for cascaded speech-LLM pipelines, jointly optimizing ASR and TTS components through reward modeling and preference learning.

PROJECTS

Hate Speech Mitigation Using LLMs — Master's Thesis, Politecnico di Torino. Developed novel methodology combining adversarial debate frameworks with RAGs to address hate speech ideologies. Implemented four-phase pipeline with adaptive mitigation strategy and cross-model evaluation, achieving 47% improvement in persuasiveness, 34% enhancement in evidence quality, and 32% toxicity reduction over established baselines. Paper submitted to COMPSAC2026.

Federated Learning with Novel Aggregation Method — Course Project, Politecnico di Torino. Developed FedAvg2Rep, a novel federated learning algorithm with secondary aggregation to improve global model performance. Evaluated on CIFAR-100 and Shakespeare datasets across 100+ distributed clients under non-IID conditions. [GitHub]

Adversarial Defense in Vision Transformers — Course Project, Politecnico di Torino. Built a concept-based adversarial detection framework for Vision Transformers, achieving 92.2% true positive rate and 84.2% AUC score against FGSM attacks. Developed MLP-based concept predictors monitoring semantic consistency in ViT attention mechanisms, detecting adversarial perturbations with 74.5% accuracy while maintaining 33.1% false positive rate across multiple epsilon values. [GitHub]

SKILLS

Programming Languages	Python (Advanced), Java, C++, C, CUDA (Intermediate), pySpark
ML & DL Frameworks	PyTorch, TensorFlow, scikit-learn, Hugging Face Transformers
Specializations	NLP, LLMs, Computer Vision, Speech Processing, RLHF, Adversarial ML
Data & Analysis	NumPy, Pandas, Matplotlib
Development Tools	Git, GitHub, Jupyter Notebooks, Linux/Unix, Shell Scripting, HPC