Project Title: A Multi-Agent LLM Framework for Detecting Narrative Attacks in Social and News Media

Project Description: This internship will develop a Multi-Agent LLM framework to detect brand narrative attacks (coordinated or organic misinformation narratives, rumor cascades, out-of-context media, impersonation, and claim variants) across social and news media. The intern will build an end-to-end pipeline where LLM agents collaborate to: (i) ingest and normalize multi-source content, (ii) extract and canonicalize claims, (iii) cluster content into evolving narratives, (iv) classify attack types and severity, (v) retrieve and verify evidence from credible sources, and (vi) generate citation-grounded incident briefs for rapid triage. The work will emphasize factual accuracy and robustness, minimizing hallucinations via retrieval-first workflows, and consistency checks.

Project Type: Research and development

Duties/Activities: The students will be asked to work on one or more of the following:

- Design an LLM-agent workflow (planner + retriever + verifier + summarizer + auditor agents) for narrative attack detection and reporting
- Implement attack taxonomy classification (e.g., fabricated quote, out-of-context media, conspiracy framing)
- Integrate retrieval and evidence verification (source ranking, cross-source consistency checks, citation constraints)
- Evaluate the system

Required Skills:

- Fundamental knowledge of **AI / Machine Learning / NLP**
- **Programming proficiency in Python** (data processing, model inference, pipelines)
- **Strong problem-solving and research ability:** analyzing approaches, implementing prototypes, working with large-scale text data

Preferred Intern Academic Level: PhD, MSc, senior undergraduate students enrolled in CS/CSE (or related fields.

Learning Opportunities: You will gain hands-on experience building agentic LLM systems for misinformation, including narrative mining, retrieval-grounded verification, and rigorous evaluation, along with practical deployment and API skills. Strong contributions may lead to a research publication based on the framework, dataset, and results.

Expected Team Size: 2-3 people

Mentors

Firoj Alam (https://firojalam.one/)
Ali Shahroor
Mohamed Kmainasi