Trustworthy Models and Data

Presented By: Sabbir Hossain Ujjal

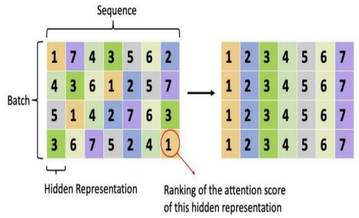
INFINITY LAB
Prof. K.P. Subbalakshmi
ksubbala@stevens.edu; https://sites.google.com/stevens.edu/infinitylab/home

Trustworthy Machine Learning

- We work on several aspects of trustworthy machine learning – at the fundamental and application level
- Our major thrusts include
  - Information theoretic approaches to explainable AI
  - Large Language Model(LLM) vulnerabilities such as hallucination, bias, etc.
  - Applications to misinformation detection

Foundational Questions

- **Can Attention Values be Used as Explanation: An Information Theoretic Perspective:**
  - In **Revisiting Attention Weights as Explanations from an Information Theoretic Perspective, NeurIPS W, 2022**[1], we show that some kinds of attention mechanisms can, under some circumstances behave as proxies for explanations.



- **Causal-TGAN: Generating tabular data using underlying causal relationships :**
  - Synthetic data generation is an important solution to privacy leakage and data shortage
  - Most generative models ignore causal forces at play between different data points
  - In **Causal-TGAN: Modeling Tabular Data Using Causally-Aware GAN, ICLR W, 2022**[2], we propose a method to capture causal relations in generating tabular data.

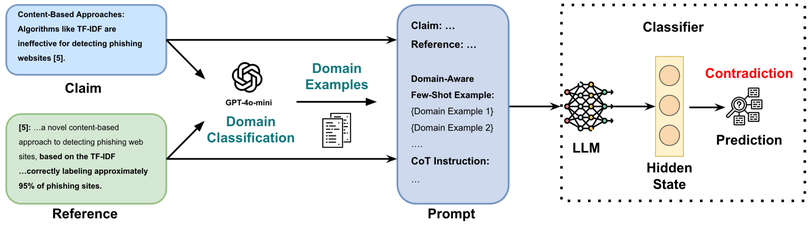
LLM Vulnerabilities : Hallucination Detection from RAG System's Output

Goal: To understand and reduce issues like hallucination and bias in large language models (LLMs) which can significantly impact the reliability and fairness of AI-driven decisions.

Our Research:

- We analyze the underlying causes of hallucination in large language models (LLMs) and propose techniques to mitigate the risk of hallucination and generating inaccurate information. Our study explores diverse approaches, including natural language reasoning, logit-level and language-level analyses, as well as information-theoretic methods, to enhance the reliability and factual consistency of LLM outputs.

- **Detecting Hallucinations in Scientific Claims by Combining Prompting Strategies and Internal State Classification**[3]
  - Domain-aware fewshot examples + Chain-of-Thought prompting + Hidden States Classification.



Model & Prompt	Score
Subtask 1	
Llama-3.1-70B-Inst, Few-Shot Prompt 2	0.49
Llama-3.3-70B-Inst, Domain-Aware Few-Shot	0.55
Llama-3.3-70B-Inst, Domain-Aware Few-Shot + CoT	0.54
Llama-3.1-70B-Inst, Few-Shot Prompt 2 + Log-Reg on hidd-stat	**0.59**
Llama-3.1-70B-Inst, Domain-Aware Few-Shot + Log-Reg on hidd-stat	**0.59**
Subtask 2	
Llama-3.1-70B-Inst, Few-Shot Prompt 2	0.40
Llama-3.1-70B-Inst, Few-Shot Prompt 2 + checklist	0.47
Llama-3.1-70B-Inst, Few-Shot Prompt 2 + Log-Reg on hidd-stat	**0.51**

Application: Misinformation Detection

Goal: Combat the spread of false narratives, hoaxes, and manipulated facts that can have significant social, political, and public health impacts. We propose natural language processing techniques to analyze content, context, and source credibility.

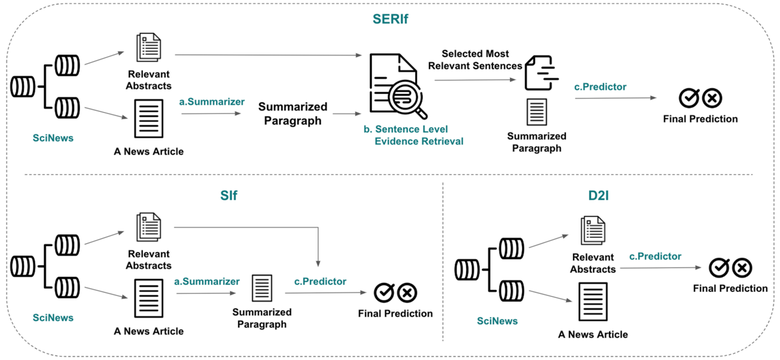
Previous Research:

- **Explainable Rumor Detection using Inter and Intra-feature Attention Networks.**[4]

- **MMCoVaR: Multimodal COVID-19 Vaccine Focused Data Repository for Fake News Detection and a Baseline Architecture for Classification.**[5]

Recent Research:

- **CoSMis: A Hybrid Human-LLM COVID Related Scientific Misinformation Dataset and LLM pipelines for Detecting Scientific Misinformation in the Wild**[6]
  - We Proposed Dimensions of Scientific Validity (DoV) guided Chain-of-Thought (CoT), can guide large language models to provide rationales for their judgments in the task of scientific misinformation detection.



STEVENS INSTITUTE OF TECHNOLOGY

[1] Wen, B., Subbalakshmi, K. P., & Yang, F. (2022). Revisiting attention weights as explanations from an information theoretic perspective. In NeurIPS 2022 Workshop on All Things Attention: Bridging Different Perspectives on Attention.
[2] Wen, B., Cao, Y., Yang, F., Subbalakshmi, K. S., & Chandramouli, R. (2022). Causal-TGAN: Modeling Tabular Data Using Causally-Aware GAN. In *ICLR Workshop on Deep Generative Models for Highly Structured Data.*
[3] Cao, Y., Yu, C.-N., & Subbalakshmi, K. P. (2025). Detecting Hallucinations in Scientific Claims by Combining Prompting Strategies and Internal State Classification. *Proceedings of the Fifth Workshop on SDP.*
[4] Chen, M., Wang, N., & Subbalakshmi, K. P. (2020). Explainable rumor detection using inter and intra-feature attention networks. In *TrueFact KDD Workshop*
[5] Chen, M., Chu, X., & Subbalakshmi, K. P. (2022). MMCoVaR: Multimodal COVID-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In Proceedings of the 2021 International Conference on Advances in Social Networks Analysis and Mining
[6] Cao, Y., Nair, A., Jamalipour Soofi, N., Eyimife, E., & Subbalakshmi, K. (2025). CoSMis: A hybrid human-LLM COVID related scientific misinformation dataset and LLM pipelines for detecting scientific misinformation in the wild. In AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation (PDLM)