A Mechanistic Approach for Detecting Biases, Threats, and Vulnerabilities in Large Language Models

– BENGAL Proposers' Day Lightning Talk –

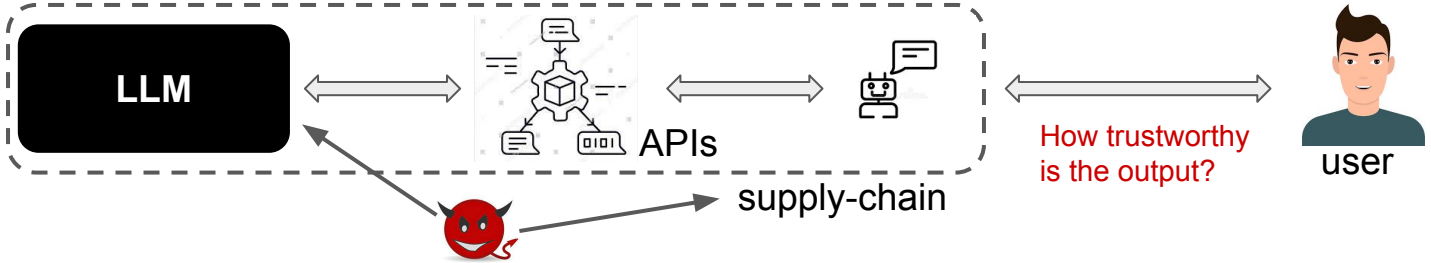
N. Benjamin Erichson
erichson@icsi.berkeley.edu

International Computer Science Institute, an Affiliated Institute of the UC Berkeley

October 24, 2023

Motivation

- LLMs are increasingly being deployed into various real-world applications.

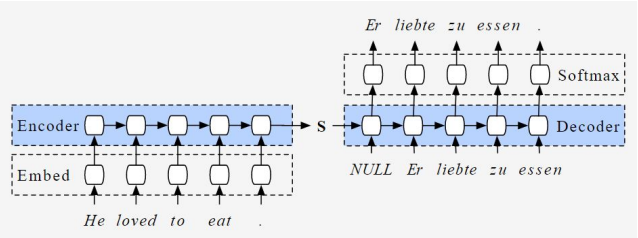


- **Intrinsic Biases:**
  - LLMs might make up facts (“hallucinate”), or generate polarized content.
  - LLMs also have the tendency to reproduce biases, hate speech, or stereotypes.
  - **Why?** LLMs are pretrained on untrusted datasets, which potentially include social biases, etc.
- **Vulnerabilities**:
  - LLMs are prone, among others, to prompt injection (PI) attacks, and backdoor attacks.

⟶ Standard accuracy metrics do not measure truthfulness, and thus are of limited use.

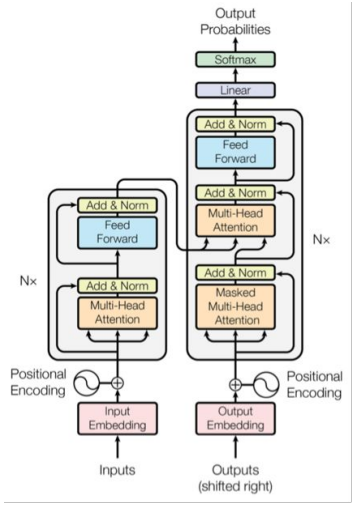
Problem Class and Threat Modes

- We focus on **large sequence-to-sequence models** for summarization, translation, and dialog generation. Models of interest include recurrent neural networks and encoder-decoder Transformer architectures.



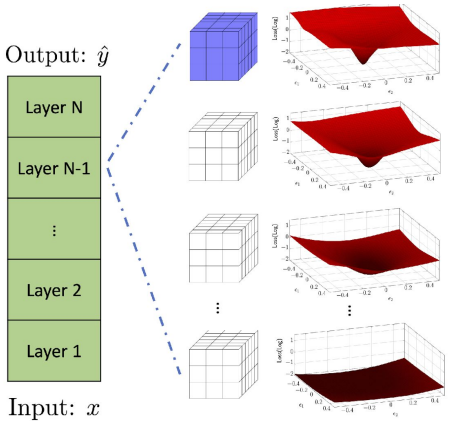
- **Threat Mode 1: Biased and Manipulated Content.**
  - LLMs have the tendency to reproduce biases.
  - LLMs can be prompted to provide wrong summaries, propagate disinformation, or hide specific facts.

- **Threat Mode 2: Intrusion.**
  - LLMs might contain natural and/or artificial backdoors that allow an adversary to trigger a malicious protocol or gain different levels of access to the LLM.



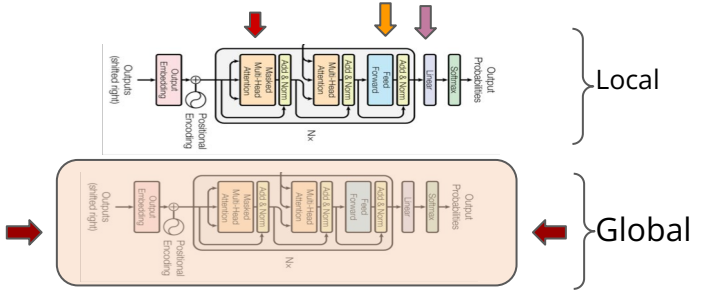
Capabilities

We propose a mechanistic approach to understand and detect biases, threats and vulnerabilities in large language models.



Local understanding: We propose local techniques such as weight analysis (e.g., eigen analysis, hessian analysis, weight statistics) to understand and analyze layers and to correlate the weight signals with biases, and vulnerabilities.

- This approach is, for instance, highly effective for backdoor detection in language models, but also applicable to other metrics of interest.
- Weight analysis also allows us to track behavior during training.



Global understanding: We propose global techniques such as input-output analysis (e.g., prompt generation, noise-response analysis) to understand the conditions under which the model may fail.

We will construct reference models by utilizing unlearning techniques, compare behavior between different LLMs, and at different training phases.

Our **capabilities**: (i) adversarial machine learning, (ii) language model probing, (iii) prompt injection.