Identifying LLM Information Leaks: Using Compositionality to Detect Anomalies and Identify Source in Generated and Human-made Data

Abstract

Anomaly and leak detection are important devices for ensuring users are not exposed to unauthorized information and the overall security of LLMs. The study of Data Loss Prevention (DLP) and Anomaly Detection (AD) encompasses the broader field used to tradition ally secure LLMs using guardrails to prevent jailbreaking and prompt-hacking. In this paper, we propose the use of language model compositionality, as proposed by SecureLLM, in conjunction with model perplexity as an additional tool for AD. Using compositional security and compositional dataset, we demonstrate that our method can effectively discriminate and determine the source composition used to generate any given output from plain text using a fine-tuned LLM. Additionally, we also show our method accurately classifies human generated samples from natural language stories, to include crossover stories that contain two or more elements from the original source. Both of these methods could be deployed to safety mechanisms for LLMs today.

Coming Soon