Protecting Your Privacy: A Constitutional Filter to Prevent AI From Inferring Your Personal Information

Du, Siyuan (School: American Heritage School of Boca Delray)

Current state-of-the-art large language models (LLMs), such as GPT, are able to accurately infer invasive details about individuals from anonymous online posts. This combined with the ease of compiling web data (such as forum posts) means privacy violations at an unprecedented scale. Using the obtained personal details (Ex. a user's age, gender, and location), malicious users can perform mass scams, blackmail, intrusive monitoring, etc. The problem worsens as LLM inference capabilities improve and usage costs decrease, meaning this will only get worse without an intervention. Despite this, there has been no research on developing a solution. This research aims to fix that by developing an easily implementable constitution-based filter that can be overlaid over any LLM. The filter works with a constitution containing principles that prohibit the inference of different categories of personal information. The filter will review LLM responses for violations of the principles. If found, the filter will requery the LLM to modify the response based on the violation. During testing, GPT-4o (advanced GPT model) had an average 84% accuracy when inferring personal information from a synthetic, text-based, dataset simulating reddit forums. The filter successfully blocked the response 99% of the time. When performing image-based inference on city images similar to those posted by real people online, GPT-4o had an accuracy of 88%, and the filter blocked 95% of the responses. This research presents an effective and implementable filter to block LLM enabled privacy violations on a large scale.

Awards Won: