CAN LLMS KEEP A SECRET? TESTING PRIVACY IMPLICATIONS OF LANGUAGE MODELS VIA CONTEXTUAL INTEGRITY THEORY

Niloofar Mireshghallah[1]* **Hyunwoo Kim**[2]*
Xuhui Zhou[3] **Yulia Tsvetkov**[1] **Maarten Sap**[2,3] **Reza Shokri**[4] **Yejin Choi**[1,2]
[1]University of Washington [2]Allen Institute for Artificial Intelligence
[3] Carnegie Mellon University [4] National University of Singapore
{niloofar,yuliat,yejin}@cs.washington.edu, {hyunwook}@allenai.org
{xuhuiz,maartensap}@cmu.edu, {reza}@comp.nus.sg.edu

ABSTRACT

The interactive use of large language models (LLMs) in AI assistants (at work, home, etc.) introduces a new set of inference-time privacy risks: LLMs are fed different types of information from multiple sources in their inputs and are expected to reason about what to share in their outputs, for what purpose and with whom, within a given context. In this work, we draw attention to the highly critical yet overlooked notion of contextual privacy by proposing CONFAIDE,[1] a benchmark designed to identify critical weaknesses in the privacy reasoning capabilities of instruction-tuned LLMs. Our experiments show that even the most capable models such as GPT-4 and ChatGPT reveal private information in contexts that humans would not, 39% and 57% of the time, respectively. This leakage persists even when we employ privacy-inducing prompts or chain-of-thought reasoning. Our work underscores the immediate need to explore novel inference-time privacy-preserving approaches, based on reasoning and theory of mind.

1 INTRODUCTION

There has been a surge of attention on privacy violations centering around the training data of large language models (LLMs), specifically with regard to personally identifiable information (e.g., social security numbers and addresses; Carlini et al. (2022)). However, LLMs are now provided with information from different sources in their inputs at inference time (Abdelnabi et al., 2023; Zhou et al., 2023b), and they need to reason about what to share in the output, for what purpose and with whom. We set out to answer the under-explored question "*Can LLMs reason about the* ***implications of contextual privacy*** *in interactive settings?*"

To study this question, we center the role of "context" in reasoning about privacy expectations, drawing from Helen Nissembaum's seminal work on "*Contextual Integrity*" theory (Nissenbaum, 2004). This theory proposes that the proper flow of information should be maintained within specific social contexts, and a privacy breach happens when the information flows against the contextual norm. For example, if your healthcare provider shares your medical history, which contains sensitive health details, with an insurance company for marketing purposes, it would be a violation of contextual integrity. In contrast, sharing the same information with other providers that are treating you would not be.

Similar to the example above, inappropriate control of information flow can lead to dire consequences when interacting with LLMs, as they have access to many of our conversations (Priyanshu et al., 2023; Duan et al., 2023; Edwards, 2023; Good, 2023). This introduces a new inference-time privacy threat, which existing data-centric privacy measures (e.g., data sanitization (Heider et al., 2020) and differential privacy (Abadi et al., 2016)) cannot address (Brown et al., 2022). Instead, better social reasoning capabilities, such as *theory of mind* (i.e., tracking mental states of others), become more essential as keeping track of different people's access to a piece of information and their relations is a crucial part of the context which controls the flow of that information (Colwell et al., 2016).

*Equal Contribution
[1]Website: http://confaide.github.io