of scenarios (Jiang et al., 2021; Zhou et al., 2023a; Forbes et al., 2020), while more works start to integrate social contexts into the machine morality discourse (Kim et al., 2022b; Pyatkin et al., 2023; Jin et al., 2022).

6 Conclusion and Discussion

We introduce ConfAIde to investigate the risk of *contextual privacy* leaks in LLMs. We identify new shortcomings in terms of privacy reasoning and theory of mind, demonstrating that even models that have undergone intensive RLHF training still lack reasoning on what should and should not be shared in various contexts. Finally, we explore possible mitigations, showing that straightforward measures, such as fortifying the prompt by instructing the model to maintain privacy or using chain of thought reasoning, are insufficient. Altogether our results highlight that more fundamental solutions are needed for LLMs to safely preserve privacy while deployed in real-world applications. We discuss implications of our findings and possible future directions below.

Inference-time privacy definitions Our findings also point to an existing gap in the literature, regarding privacy definitions at inference time, which can have serious consequences. For instance a recent prompt-injection attack reverse-engineered Bing Chat's initial prompt, which is a list of statements that governs how it interacts with people who use the service (Edwards, 2023). We aim to draw attention to the necessary changes in the model deployment and use pipeline, emphasizing how this new interactive application of models introduces new privacy challenges, and we only scratch the surface of possible inference time privacy concerns. There is still a plethora of risks unexplored, such as possible leakage of in-context examples to the output, and also the contention between different data modalities in the newly ubiquitous multi-modal models (Chen et al., 2023; Duan et al., 2023; Tang et al., 2023).

Need for fundamental solutions We show that effectively addressing the issues we raise is difficult, and ad hoc safeguards (e.g., privacy-inducing prompts, chain-of-thought reasoning and output filters) are insufficient to solve the core issue of contextual privacy reasoning. Prior work on limiting bias and hallucinations in LLMs (Zhou et al., 2023c) have also demonstrated that patching solutions and safeguards can be easily bypassed with malicious inputs and that there is need for fundamental and principled inference-time approaches, such as using explicit symbolic graphical representation of each character's beliefs (Sclar et al., 2023), to enhance the decision making process considering privacy and information flow.

Theory of mind for understanding privacy Inherently, privacy and secrets create a disparity in information access among individuals. Recent work demonstrates that current LLMs struggle in interactive scenarios involving information asymmetry (Kim et al., 2023). In order to enable these models to navigate complex scenarios involving privacy, it is essential for them to possess theory of mind (ToM) capabilities – i.e., tracking and understanding distinct mental states of individuals. We hope future works will further explore the intersection of ToM and contextual privacy.

Secret revealing and moral incentives While our benchmark probes models for their privacy reasoning capabilities through the theory of contextual integrity, we do not intend to be the arbiter of privacy, nor do we aim to make normative judgments on what should and should not be revealed, as such judgments can be deeply intertwined with the moral and cultural aspects of an interaction. Social psychologists have studied the moral incentives behind why people might reveal each other's secrets, as a form of punishment (Salerno & Slepian, 2022), and we encourage future work to further study such incentives in language models more extensively.

Human-AI interaction Moreover, there is another less discussed aspect of human-AI interaction: people may feel more at ease sharing information with AI models — information they would hesitate to share with other humans — believing that their disclosures will remain secure (Hart et al., 2013). This encompasses different topics, from personal matters to corporate confidential documents (Franzen, 2023; Park, 2023). We hope our benchmark paves the way for future trustworthy AI research on aligning LLMs with human privacy expectations in practice. We encourage future work to build on our benchmark and propose privacy mitigations based on contextual reasoning.