some ability to follow the flow of information. However, the high leakage of the secret in the free-form response reveals their limitation in reasoning over such knowledge and adequately controlling the flow.

To further analyze the leakage, we plot a detailed breakdown of the results, for the best performing model, GPT-4 in Figure 3. We find information regarding sexual orientation/self-harm is most/least likely to be revealed on average, and the incentives of helping others/gaining money lead to most/least leakage. Also, we observe **contention** between different contextual factors. For instance, although the model tends to not reveal information about self-harm, it opts to do so when the incentive is helping others. We provide detailed heatmaps for other models, with and without privacy-inducing prompts (i.e., without telling the model to adhere to privacy norms, which make it leak even more) in Appendix B.3.1.

4.4 Tier 4: InfoFlow-Application Results

Table 5 summarizes the results for Tier 4.[3] We provide both average and worst-case results for the leakage metric, across 10 runs. We find that all models have relatively high levels of leakage, as they tend to regurgitate the secret discussed at the beginning of the meeting. This leakage is higher in the meeting summary task compared to the personal action item generation task. We hypothesize that this could be due to the model being instructed to generate the action item specifically for person X (who is not supposed to know the secret), whereas in the summary generation the model is instructed to generate summary for all attendees, hence it isn't able to reason that X is among the attendees and that the secret should be withheld. Additionally, we also report an aggregated metric, where we consider the response erroneous if it either leaks the secret or misses an important public action item. We observe a high error rate across all models, including GPT-4. Results without the privacy instructions can be found in Appendix B.4.

Figure 4 shows a breakdown of Tier 4 results for GPT-4, which is the best performing model in the action item generation task. Our most noteworthy observation is the model's lack of understanding of surprises. GPT-4 consistently reveals the surprise to the person who is not supposed to know about it, even using terms such as "*attend your surprise birthday party*" in the generated action items. For health issues, on the other hand, models leak them less frequently. We provide results without direct privacy-preserving instructions, as well as results from other models in Appendix B.4.

4.5 Possible Mitigation: Chain of Thought Reasoning

In this section, we present the results for Tiers 3 & 4, but as a possible mitigation we prompt the model with chain of thought reasoning (Wei et al., 2022), i.e. we add the sequence 'Take a deep breath[4] and work on this step by step.' to the instruction provided to the model, as proposed in Yang et al. (2023). We keep the prior instructions to preserve privacy as well. Once we get the response, we feed it back to the model and ask the target question from the model again (the original questions from the tier, for instance, to provide a list of action items or to summarize the meeting notes.), and only use the final response for our evaluations (as in we do not look at the steps so leakage in the steps is not going to count as a violation). Table 6 shows the results of this experiment. We can see that for almost all tasks, using chain of thought (CoT) does not improve leakage, in fact it makes the leakage more severe, which could be due to the more detailed nature of the response. Another interesting observation is that asking the model to go step by step seems to degrade the utility of the task, in the Tier 4 task of meeting summarization, as the public information drop rate increases when using CoT.

5 Related Work

Differential Privacy (DP) for LLM Training and Inference DP provides a worst-case, context independent privacy guarantee by making models trained on datasets D and D', which differ in only one record, indistinguishable, thereby providing record-level protection. DP mechanisms have been used to train ML models and LLMs on private data, to prevent memorization and leakage of training data (Abadi et al., 2016; Li et al., 2021; Yu et al., 2021; Shi et al., 2021) or in-context examples (Panda et al., 2023; Duan et al., 2023; Tang et al., 2023).

All these works, however, focus on protecting training data, without considering context, and rely heavily on having a well-defined notion of a single record. While this is ideal for tabular data, it is extremely

[3] For this tier, we drop Flan-UL2 as it struggles with the long convoluted scenarios and generates nonsensical outputs.

[4] Yang et al. (2023) find this prompt to be most effective, through a prompt optimization method.