(3) Understanding private information sharing: We evaluate the model whether it understands the concept of shared private information. Instead of explicitly specifying the information piece, we ask the model with the following prompt: "*List all the characters who shared the private information of X*". Similarly, models should not include Z while including Y in their response.

(4) Control question: To disentangle the intricate nuances from to core decision-making process, we simplify the scenarios and formulate them into binary decision questions: "*Is it ok for Y to share information about {secret}, with Z, in order to {incentive}?*". The model should always answer 'no' for these questions. The gap in performance between this task and response generation task illustrates the effect of the added nuances.

3.4 Tier 4: InfoFlow-Application

Design In this final tier, we aim to further test the privacy reasoning of LLMs in two real-world application-inspired scenarios: automatic action-item and summary generation from meeting transcripts (Ajam, 2023). As LLMs can generate more natural conversations than crowdsourcing (Kim et al., 2022a), we synthesize the meeting transcripts using GPT-4 with few-shot samples. The meeting begins with three people discussing a sensitive topic (i.e., *secret*) concerning a fourth individual, X. The attendees make it clear that X should remain unaware of this discussion. In addition, they also share an important information that everyone should be made aware of (i.e., *public information*). As the meeting progresses, X and another person join, bringing the total participants to five. Notably, neither the secret nor the public information is mentioned again. We generate 20 transcripts also in a factorial manner, by iterating through five secrets (e.g., health concerns and surprise birthday party) and four public information types (e.g., specific deadline dates and locations of group dinner). A sample meeting transcript is in Table 7, and more details can be found in the Appendix A.4.

Evaluation (1) The first task is creating a list of action items for X based on the full meeting transcript. (2) The second task is to generate a meeting summary that will be automatically sent to all attendees. In performing these tasks, it is important to *exclude the sensitive information* from the earlier discussion related to X, while *including the vital public information*. This presents a trade-off between privacy and utility. If the model is overly conservative about sharing information, it might inadvertently omit the crucial public information along with the confidential information. We ask the model task-specific prompts including privacy-preserving instructions, which can be found in Appendix A.4. For both tasks, we use exact string-match to detect the sensitive information and public information included in the model outputs.

3.5 Human Annotations

We collect human expectations and preferences for tiers 1 through 3 using Amazon Mechanical Turk (MTurk). We ask five workers for each sample. In tiers 1 and 2, we follow Martin & Nissenbaum (2016), asking workers for their individual opinions on the sample and taking the average. For tier 3, we present workers with a choice task between two sample responses: one that reveals X's secret and another generic response that omits any mention of X's secret. We then determine the preferred response based on the majority vote from the five workers

Results For tiers 1 and 2, we find our results to be closely aligned with the initial results of Martin & Nissenbaum (2016), demonstrating a correlation of 0.85, overall. In tier 3, out of 270 scenarios, only 9 received a majority vote to disclose private information, and each of them received no more than 3 out of 5 votes. Meanwhile, 90% of the samples that preferred to keep the information private received at least 4 votes. The pair-wise agreement for tiers 1, 2.a, 2.b, and 3 are 70.7%, 76.9%, 74.6%, and 90.8%, respectively. More details can be found in the Appendix B.1.

4 Experimental Results

In this section, we first provide a summary of results in terms of model alignment with human judgments, and then discuss a more detailed tier-by-tier analysis. We run our experiments on the following models: GPT-4, ChatGPT, InstructGPT[2], Llama-2 Chat (70B), Llama 2 (70B) and Flan-UL2 (OpenAI, 2023; 2022; Ouyang et al., 2022; Touvron et al., 2023; Tay et al., 2022). We report our metrics averaged over 10 runs.

[2] `gpt-4-0613`, `gpt-3.5-turbo-0613`, `text-davinci-003`