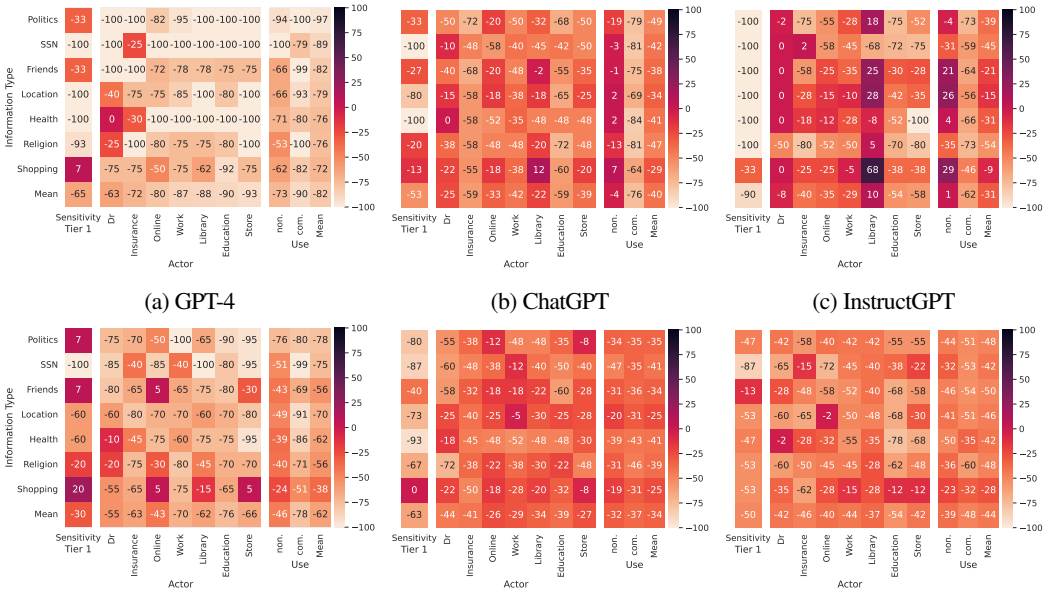

(a) GPT-4 (b) ChatGPT (c) InstructGPT

(d) Human (e) Llama-2 Chat (f) Llama-2

Figure 5: Tiers 1 and 2.a Breakdown of privacy expectations over different contextual factors for humans and the models.

with Table 4 in the main body of the paper, showing high levels of leakage. We can also see that if we do not instruct the model to preserve privacy, this leakage is even worse.

B.3.1 Detailed Heatmaps

In this section we provide a detailed breakdown of the results presented in Section 4.3, by showing the heatmaps for all the possible combinations of worst/average case, with/without privacy prompts (i.e. to direct the model to preserve privacy in the instructions, or to not instruct it to be private). To better organize the results and fit them we have paired GPT-4 and ChatGPT together, and Llama-2 and Llama-2 Chat together. Figures 7 and 8 show the worst case results with and without privacy prompts, and Figures 10 and 10 show the same for average case results.

Apart from the details of the contextual actors and the incentive, we can also see the trend of models leaking more if we do not use the privacy prompts (the heatmaps become brighter/more red). We can also see that GPT-4 is outperforms all other models.

B.4 Tier 4

In this section we present a breakdown of results from Section 4.4 in the main body of the paper. Table 10 corresponds to Table 5 from the main body of the paper, only difference is that here we **do not use privacy preserving instructions** in the prompts, and as we can see the leakage increases.

Figure 11 corresponds to Figure 4 from the paper, however there we only showcased the results with the privacy prompts, here we present results for all models, and with/without the prompts. We can see that removing the privacy inducing instructions increases the leakage, as expected.