Table 1: Pearson's correlation between human and model judgments for each tier, higher values show more agreement. We see the correlation decrease as we progress through tiers and tasks become more nuanced.

Tier	GPT-4	ChatGPT	InstructGPT	Llama-2 Chat	Llama-2	Flan-UL2
Tier 1: Info-Sensitivity	0.86	**0.92**	0.49	0.71	0.67	0.71
Tier 2.a: InfoFlow-Expectation	0.47	0.49	0.40	0.28	0.16	**0.50**
Tier 2.b: InfoFlow-Expectation	**0.76**	0.74	0.75	0.63	-0.03	0.63
Tier 3: InfoFlow-Control	**0.10**	0.05	0.04	0.01	0.02	-0.18

Table 2: Value of sensitivity scores (Tier 1) and privacy expectations for information flow (Tier 2), averaged over all the samples in each tier. Lower values indicate less willingness to share information. We find models' conservativeness decreases on average, as we progress through tiers.

Metric	Human	GPT-4	ChatGPT	InstructGPT	Llama-2 Chat	Llama-2	Flan-UL2
Tier 1: Info-Sensitivity	-29.52	-64.76	-53.33	**-90.48**	-62.86	-50.48	-53.33
Tier 2.a: InfoFlow-Expectation	-62.04	**-81.73**	-39.90	-30.51	-34.23	-43.52	-43.52
Tier 2.b: InfoFlow-Expectation	-39.69	**-57.65**	-21.43	11.02	-2.09	-42.55	-41.28

Table 3: Information type and contexts in which the model vs. human judgment gap on privacy expectations is the largest, with the model being much more/less conservative (Most/Least conservative rows). Each table slot shows Information Type/Actor/Use, with NC being non-commercial and $ being commercial use.

	Model v. Human	GPT-4	ChatGPT	InstructGPT	Llama-2 Chat	Llama-2	Flan-UL2
T 1	Most Conservative	Religion	Politics	Friends	Politics	Shopping	Friends
	Least Conservative	SSN	SSN	SSN	SSN	SSN	Shopping
T 2.a	Most Conservative	SSN/Insurance/NC	SSN/Insurance/NC	SSN/Insurance/NC	Health/Dr/NC	Health/Dr/NC	SSN/Insurance/NC
	Least Conservative	SSN/Insurance/$	SSN/Dr/NC	SSN/Insurance/$	Location/Work/$	Health/Dr/$	SSN/Insurance/$
T 2.b	Most conservative	Shopping/Online/NC	Religion/Work/NC	Religion/Dr/$	Friends/Library/NC	Health/Dr/NC	Shopping/Online/NC
	Least Conservative	Shopping/Education/NC	Health/Library/NC	Politics/Insurance/NC	Politics/Insurance/NC	Health/Insurance/$	Health/Library/NC

4.1 All Tiers: Alignment with Human Judgement

Table 1 reports the correlation between human and model judgments, using the Pearson correlation score (see § 3.5 for annotation details). For Tier 4, since we build situations where the AI agent must not reveal private information, we do not collect human annotations, and only report error rates in § 4.4. We observe two main trends in the table: (1) **As we move up the tiers, the agreement between humans and the models decreases**, and (2) models that have undergone heavy RLHF training and instruction tuning (e.g. GPT-4 and ChatGPT) tend to align more closely with human judgment. Nevertheless, an alarming gap still exists for the higher tiers, pointing to potential issues for more complicated tasks. We dive deeper into these issues in the sections below.

4.2 Tiers 1-2: Info-Sensitivity and InfoFlow-Expectation Results

Table 2 shows the average sensitivity score over information types (Tier 1) and average privacy expectation scores for the factorial combination of 'information type', 'actor' and 'use' (see § 3.3, Tier 2). Lower scores indicate a lower willingness to share the information, denoting greater conservativeness. In Tier 1, all models are more conservative than humans, with InstructGPT being the most conservative on average. Moving on to Tier 2.a, all models, except GPT-4, show decreased conservativeness. In contrast, GPT-4's conservativeness increases on average, which is similar to the human judgments. Finally, in tier 2.b, we see even less conservativeness on average, with InstructGPT showing the highest surge.

To better understand the contexts regarding models' conservativeness, we provide a breakdown in Table 3. This table shows the information types/contexts where the absolute difference between human and model judgments are the largest (i.e., most/least conservative). The most surprising result is in Tier2.a, concerning SSN. For example, we find GPT-4 is much more conservative than humans when it comes to sharing SSN with insurance for a non-commercial purpose (i.e., to detect fraud). Conversely, it is much more permissible when the same information is shared for a commercial reason (i.e., to sell to drug stores for marketing purposes), which is alarming. These results indicate **possible misjudgments even in simple scenarios**.

Finally, to zoom in on how the progression of tiers affects a single model's judgment over different contextual factors, we plot the breakdown in Figure 2. We can see how context shapes the model's judgment, as SSN, a highly sensitive information type (Tier 1) is deemed less sensitive when it is to be shared with insurance (-100 to -25; Tier 2.a). We can also see how sharing SSN with a doctor becomes much less of a pri-