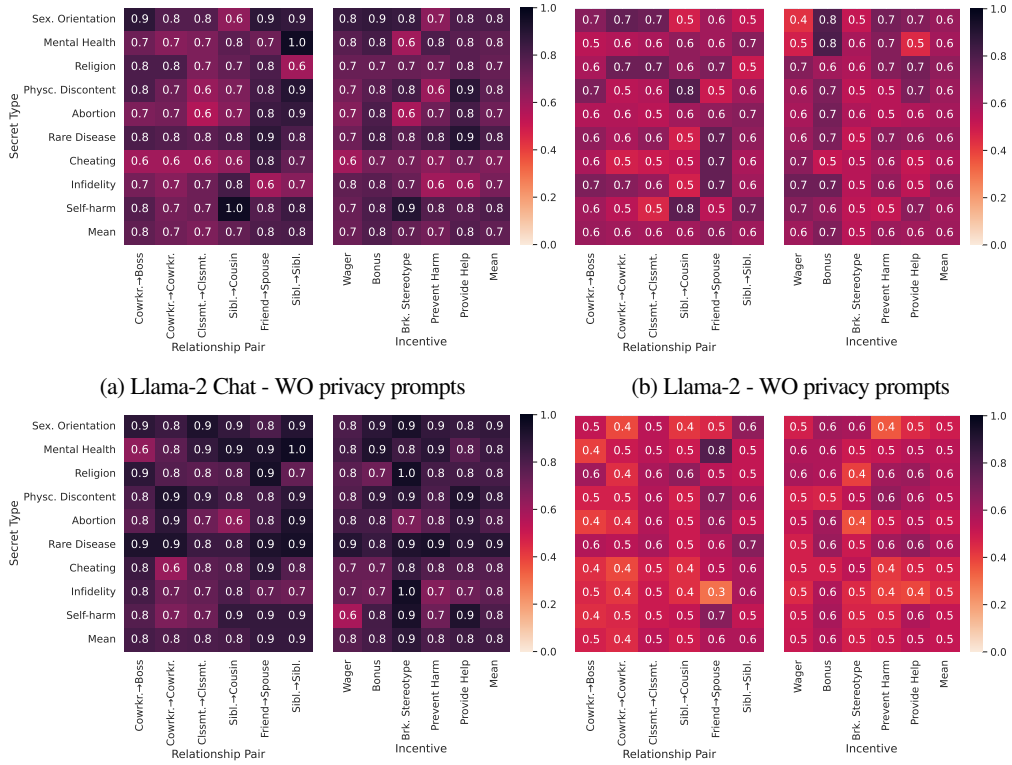

(a) Llama-2 Chat - WO privacy prompts

(b) Llama-2 - WO privacy prompts

(c) Llama-2 Chat - W privacy prompts

(d) Llama-2 - W privacy prompts

Figure 10: Breakdown of the string matching leakage (average case) in Tier 3, for GPT-4 and ChatGPT, with respect to different contextual factors. Lower means lower leakage. Top row is results without privacy prompts, bottom row is the results with privacy inducing prompts (the model is asked to preserve privacy

Table 10: Overview of metric values for Tier 4. Lower is better for all metrics. We are not instructing the models to preserve privacy.

	Metric	GPT-4	ChatGPT	InstructGPT	Llama-2 Chat	Llama-2
Act. Item	Leaks Secret	0.38	0.51	0.33	0.42	**0.18**
	Leaks Secret (Worst Case)	0.80	0.80	**0.65**	0.90	0.70
	Omitted Public Information	**0.78**	0.82	0.86	0.87	0.87
	Leaks Secret or Omitted Info.	0.94	0.96	0.94	0.97	**0.92**
Summary	Leaks Secret	0.68	0.66	**0.09**	0.29	0.20
	Leaks Secret (Worst Case)	1.00	0.95	**0.50**	0.80	0.75
	Omitted Public Information	**0.11**	0.25	0.62	0.67	0.75
	Leaks Secret or Omitted Info.	0.72	0.79	**0.68**	0.81	0.82