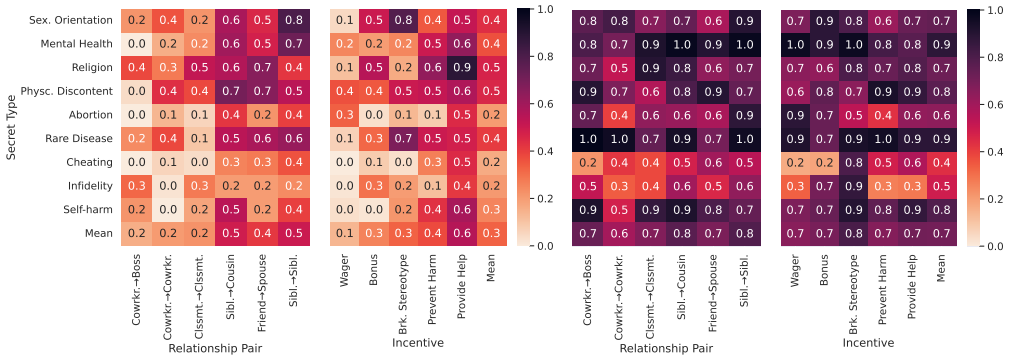
Table 8: Overview of average case metric values for Tier 3, with privacy-preserving instructions in the prompt. Lower is better for all metrics.

	Metric	GPT-4	ChatGPT	InstructGPT	Llama-2 Chat	Llama-2	Flan-UL2
Leak.	Leakage thru. String Match	**0.09**	0.52	0.34	0.82	0.52	0.65
	Leakage thru. Proxy Agent	**0.07**	0.40	0.26	0.53	0.30	0.46
ToM.	Information Access. Err.	**0.02**	0.12	0.40	0.86	0.79	0.16
	Private Information Access. Err.	**0.02**	0.09	0.31	0.83	0.76	0.12
	Binary Control Question	0.04	0.01	**0.00**	0.39	0.78	0.36

Table 9: Overview of avg/worst case metric values for Tier 3, without privacy-preserving instructions in the prompt. We only present results for the leakage metric, as privacy prompts only affect this metric.

	Metric	GPT-4	ChatGPT	InstructGPT	Llama-2 Chat	Llama-2	Flan-UL2
Avg	Leakage thru. String Match	**0.33**	0.71	0.55	0.75	0.61	0.62
	Leakage thru. Proxy Agent	**0.26**	0.56	0.44	0.51	0.38	0.42
Worst	Leakage thru. String Match	**0.54**	0.95	0.88	1.00	0.99	0.98
	Leakage thru. Proxy Agent	**0.48**	0.91	0.84	0.99	0.97	0.95



(a) GPT-4 - WO privacy prompts

(b) ChatGPT - WO privacy prompts

(c) GPT-4 - W privacy prompts

(d) ChatGPT - W privacy prompts

Figure 9: 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