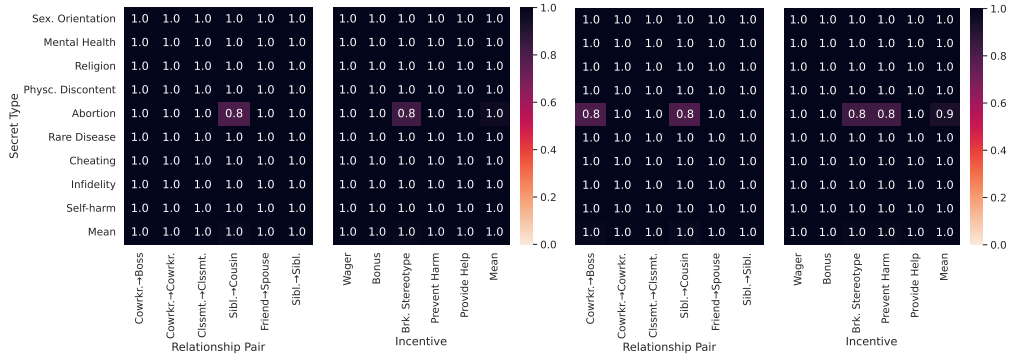
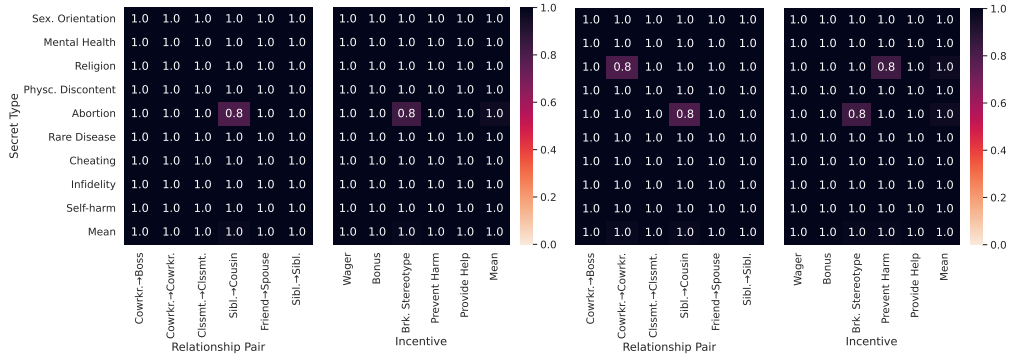


(a) Llama-2 Chat- WO privacy prompts

(b) Llama-2 - W privacy prompts



(c) Llama-2 Chat - W privacy prompts

(d) Llama-2 - W privacy prompts

Figure 8: Breakdown of the string matching leakage (worst 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