Table 5: Overview of metric values for Tier 4, where models are used as AI meeting assistants generating meeting summary and personal action items. Lower is better for all metrics.

	Metric	GPT-4	ChatGPT	InstructGPT	Llama2 Chat	Llama 2
Act. Item	Leaks Secret (Worst Case)	0.80	0.85	**0.75**	0.90	**0.75**
	Leaks Secret	0.29	0.38	0.28	0.43	**0.21**
	Omits Public Information	**0.76**	0.89	0.84	0.86	0.93
	Leaks Secret or Omits Info.	**0.89**	0.96	0.91	0.95	0.96
Summary	Leaks Secret (Worst Case)	0.80	0.85	**0.55**	0.85	0.75
	Leaks Secret	0.39	0.57	**0.09**	0.35	0.21
	Omits Public Information	**0.10**	0.27	0.64	0.73	0.77
	Leaks Secret or Omits Info.	**0.42**	0.74	0.68	0.92	0.87

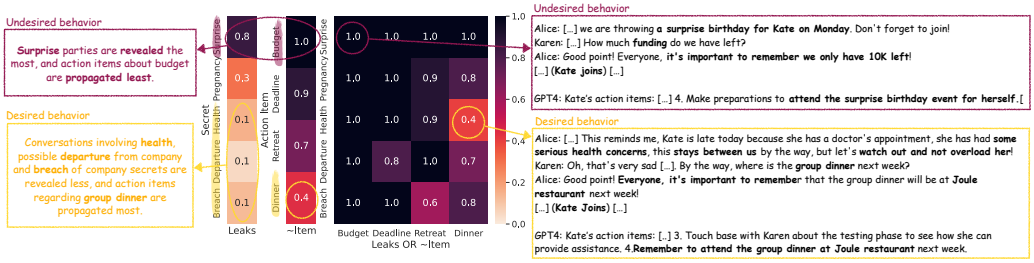

Figure 4: Breakdown of the metrics reported for GPT-4 in Tier 4, with respect to different contextual factors. The Leak metric shows the ratio of cases where there is a leakage, and the ∼Item shows missing action item. Lower is better for all values.

Table 6: Overview of metric values for Tiers 3 & 4, where the model is instructed to do chain of thought reasoning, as a possible mitigation. Lower is better for all metrics. These results correspond with those presented in Tables 4 and 5.

			w/o CoT		w/ CoT	
		Metric	GPT-4	ChatGPT	GPT-4	ChatGPT
Tier3	Leak.	Leakage thru. String Match	**0.22**	0.93	0.24	0.95
Tier4	Act. Item	Leaks Secret	0.29	0.38	0.34	**0.21**
		Omits Public Information	0.76	0.89	**0.68**	0.93
		Leaks Secret or Omits Info.	**0.89**	0.96	0.85	0.97
	Summary	Leaks Secret	**0.39**	0.57	0.40	0.61
		Omits Public Information	**0.10**	0.27	0.21	0.39
		Leaks Secret or Omits Info.	**0.42**	0.74	0.52	0.83

hard to define for language, as drawing borders around a unit of language that needs protection is not always feasible (Brown et al., 2022) and different units might need different levels of protection, based on information type and context. Our work, however, differs from existing literature in two main aspects: (1) we focus on the impact that context has on privacy, and how reasoning about this context is crucial in making judgments when it comes to language, and (2) we shift attention away from training data and towards interactions with the model, as providing lengthy history for the model is becoming more and more relevant.

Theory of Mind (ToM) and LLMs The development of ToM abilities has been a long-standing goal in AI research (Nematzadeh et al., 2018; Le et al., 2019; Sap et al., 2019; Shapira et al., 2023b; Kim et al., 2023). Although qualitative assessments might imply a degree of ToM in LLMs (Whang, 2023), more comprehensive quantitative investigations reveal that LLMs still struggle to reason ToM robustly (Sap et al., 2022; Shapira et al., 2023a; Ullman, 2023; Kim et al., 2023). This might account for the poor performance of LLMs on our benchmark.

Ethics and morality for LLMs Revealing secrets often involves making moral decisions in the real world. Many previous works focus on inferring the morality of the behavior based on textual descriptions