various social domains or "contexts" (e.g. health, work, family, civil and political, etc). Privacy violations occur when information flows deviate from the established norms and principles of the particular context. For example, information being shared inappropriately or without consent, or being used for purposes that were not intended within that context. Conversely, appropriate information flows are those that conform with contextual information norms, such as sharing news on social media, or income information with the IRS (Martin & Nissenbaum, 2016).

Contextual integrity singles out five critical parameters to describe data transfer operation. Assessing the privacy impact of information flows requires the values of all five parameters to be specified (Nissenbaum, 2018): *data subject* (e.g. patient, shopper), *sender and receiver of the data* (e.g. hospital, bank), *information type* (e.g. medical, financial) and *transmission principle* (e.g. coerced, sold). By fully specifying the contextual actors, the framework of contextual integrity provides a more expressive way for highlighting variables that are relevant to privacy. It builds on the intuition that the capacities in which actors function are crucial to the moral legitimacy (Salerno & Slepian, 2022) of certain flows of information. This holds true even when it might appear that it does not: when people remark that certain information is 'secret' what they usually mean it is secret in relation to some actors, rather than absolutely (Nissenbaum, 2009).

**Theory of Mind and Secret Keeping** Assessing the appropriateness of information flow and privacy (i.e., contextual integrity) relies heavily on *identifying the context* and *reasoning* over social norms along with the possible consequences of sharing vs. not sharing (Kökciyan, 2016; Shvartzshnaider et al., 2019; Solove, 2023). Theory of mind (ToM) — the ability to comprehend and track the mental states and knowledge of others (Premack & Woodruff, 1978) — plays a crucial role in discussions surrounding privacy in context and the act of keeping secrets (Colwell et al., 2016). This not only shapes the *context* when making privacy-related decisions, but also helps us to predict how others will react and recognize the consequences of our actions (Byom & Mutlu, 2013; Colwell et al., 2016; Slepian, 2022). In our work, we leverage theory of mind related tasks to evaluate how models manage secrets and to craft more intricate scenarios within our benchmark (Li et al., 2017).

## 3 CONFAIDE: BENCHMARKING CONTEXTUAL PRIVACY REASONING IN LLMS

In this section, we introduce the design and evaluation methods of CONFAIDE. Specifically, we aim to assess the contextual reasoning abilities of large language models (LLMs) in terms of information flow and privacy. CONFAIDE consists of four tiers, each having distinct evaluation tasks. The design of the first two tiers draws inspiration from Martin & Nissenbaum (2016), which is a legal study empirically measuring people's privacy expectations by operationalizing the theory of *contextual integrity* (see §2). As we progress to the higher tiers, the contexts become more intricate, necessitating a more pronounced involvement of social reasoning, such as *theory of mind*. Samples of our benchmark are in Figure 1.

### 3.1 TIER 1: INFO-SENSITIVITY

**Design** We first aim to assess LLMs on their basic ability to understand the sensitivity of a given information. We follow the approach of Martin & Nissenbaum (2016) by providing certain *information types* and asking how sensitive they are, without any further context. To avoid anthropomorphising LLMs, we frame questions in terms of how people might perceive the sensitivity of the information, rather than asking the model for its own opinion. Specifically, we use the ten information types taken from Pew's study (Madden, 2014), such as social security number, health status, and messages. More details are in Appendix A.1.

**Evaluation** The model is asked to rate each of their sensitivity with the following prompt that has four rating options: "*Please indicate how sensitive people would consider the following information to be — even if some people and organizations already have access to it. Do so by choosing one of the following ratings: 4) very sensitive, 3) somewhat sensitive, 2) not too sensitive, 1) not at all sensitive.\nInformation: {information type}\nAnswer:*". This prompt is taken directly from Martin & Nissenbaum (2016). We compare the model's responses with human annotations (§3.5).

### 3.2 TIER 2: INFOFLOW-EXPECTATION

**Design** Next, we evaluate the models' expectations of information flow using the *factorial vignette survey* design from Martin & Nissenbaum (2016). The vignettes are systematically constructed situation templates