Personalizing Agent Privacy Decisions via Logical Entailment

James Flemings*
University of Southern California
California, USA
jamesf17@usc.edu

Ren Yi
Google Research
New York, USA
ryi@google.com

Octavian Suciu
Google Research
New York, USA
osuciu@google.com

Kassem Fawaz†
University of Wisconsin - Madison
Wisconsin, USA
kfawaz@wisc.edu

Murali Annavaram
University of Southern California
California, USA
annavara@usc.edu

Marco Gruteser
Google Research
New York, USA
gruteser@google.com

Abstract

Personal large language model (LLM) agents increasingly perform tasks that require access to user data, raising concerns about appropriate data disclosure. We show that relying solely on LLMs to make data-sharing decisions is insufficient. Prompting LLMs to ground their decisions on contextual privacy norms fails to capture individual users' privacy preferences, while providing prior user data-sharing decisions through in-context learning (ICL) leads to unreliable and opaque reasoning. To address these limitations, we propose ARIEL (Agentic Reasoning with Individualized Entailment Logic), a framework that combines LLMs with rule-based logic to enable structured, personalized privacy reasoning. The core mechanism of ARIEL determines whether a user's prior decision on a data-sharing request *logically entails* the same decision for a new request. Experimental evaluations using advanced models and public datasets show that ARIEL reduces the F1 error rate for appropriate judgments by **40.6%** compared to standard ICL-based reasoning, indicating that ARIEL is effective at correctly judging requests where the user would approve data sharing. These results demonstrate that integrating LLMs with logical entailment provides an effective and interpretable approach for automating personalized privacy decisions.

Keywords

privacy, agents, large language models, personalization

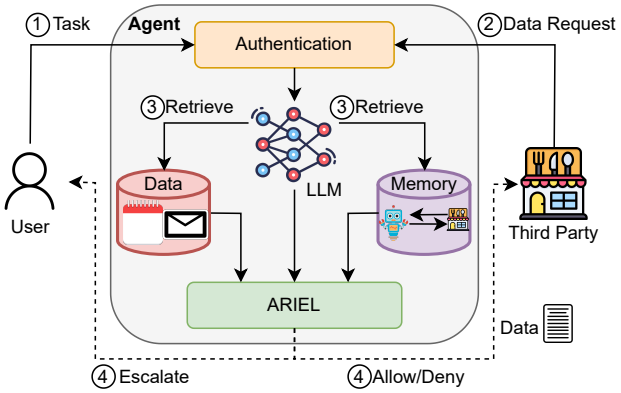

Figure 1: A schematic visualization of the problem setup with ARIEL, our proposed privacy-reasoning framework, operating in a larger end-to-end agentic system. ① A user sends a task to the agent to complete. ② A third party requests user data in order to complete the task. ③ After the data request has been authenticated, the agent retrieves the relevant user data and prior user judgments on data-sharing requests from memory. ④ The agent decides whether it is appropriate to share the user's information to fulfill the third party's request. ARIEL either allows or denies transmitting the user data, or escalates the data-sharing request to the user for further assistance.

1 Introduction

As personal AI agents increasingly automate tasks, they frequently encounter scenarios requiring the transmission of private user data. Consequently, the agent must determine the appropriateness of such disclosures. Consider the restaurant booking scenario in Figure 1, where a restaurant requests the user's phone number to complete the table reservation. While the agent could seek the user's permission to transmit their phone number, frequent interruptions lead to 'privacy fatigue,' causing disengagement and ineffective data control by the user [2, 17]. To alleviate this cognitive burden, the agent should ideally be capable of autonomously judging when data sharing is appropriate.

Prior to Large Language Models (LLMs), Personal Privacy Assistants (PPAs) relied on either rule-based [73, 74] or traditional machine learning (ML) approaches [4, 10]. These systems leverage a history of prior user privacy judgments to decide whether to disclose data for incoming requests. For instance, as shown in Figure 1, if a user previously authorized sharing a phone number for a restaurant reservation, the agent can utilize this precedent to approve similar future requests. However, these methods face significant limitations: (1) rule-based approaches are domain-specific and difficult to transfer without substantial modification; and (2) ML methods require extensive user data (from many users) to achieve high performance.

The integration of LLMs into agentic workflows [62] offers a promising solution to these challenges, as their In-Context Learning

*Work done while interning at Google Research.
†Work done as a visiting faculty researcher with Google.

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license visit https://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.
Proceedings on Privacy Enhancing Technologies 2026(3), 378–407
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.56553/popets-2026-0087

(ICL) capabilities enable effective generalization even from sparse data [14]. While recent studies have evaluated privacy judgments by LLMs [42, 54], they typically ground their definition of appropriateness in general Contextual Integrity (CI) norms derived from laws or societal consensus [49, 50]. However, relying solely on general norms fails to capture the nuances of user privacy decisions [12, 26, 59, 75]. In the example above, it is generally acceptable to share a phone number for a restaurant reservation, yet a user with a restrictive history of data sharing might explicitly reject this request. General norms cannot account for individual privacy decisions, thus being a barrier in the development of personalized privacy agents. This critical gap motivates the central objective of our work.

> ***Objective:*** *Personalizing LLM-based agents to make data-sharing decisions that maximize alignment with prior user decisions, minimize user intervention, and ensure auditability.*

To fulfill this objective, we identify three requirements for designing personal LLM agents.

- *Alignment.* The agent's privacy decisions must adhere to the user's prior privacy judgments, which is important for effective personalization.
- *Interpretability.* The agent must provide interpretable and traceable reasoning for every privacy judgment, allowing the user to audit the process and maintain trust in the agent's decision-making.
- *User Agency.* The agent must preserve user agency by identifying when prior user privacy judgments are insufficient or ambiguous then escalating such requests to the user rather than forcing a judgment.

We observe that current LLM-based approaches fail to simultaneously satisfy these design requirements. While leveraging ICL on prior user privacy judgments improves personalization over general privacy norms, it does not achieve effective alignment with prior user privacy decisions due to limited logical reasoning [36] and potential for LLM hallucinations [30, 31, 76]. Furthermore, the opaque nature of these models undermines interpretability by preventing verification of the reasoning steps used to reach a privacy judgment. Because LLM-generated reasoning traces are often unfaithful to the actual decision process [11, 71], users cannot effectively audit the agent, ultimately compromising their agency.

Thus, we propose ARIEL (Agentic Reasoning with Individualized Entailment Logic), a framework that satisfies these three requirements. Our key insight is to frame personalized privacy judgments as *logical entailment*: determining whether a user's judgment on a prior request entails the same judgment for an incoming request. Rather than relying on LLMs to provide an opaque appropriateness judgment, ARIEL combines LLM-based reasoning and rule-based components for reliable and aligned judgments. Specifically, ARIEL leverages an LLM to generate ontologies that capture semantic relationships from the user's prior privacy judgments (e.g., a data sensitivity hierarchy), which are then processed by a rule-based component to provide a final judgment. This framework ensures that the agent acts only when a decision can be inferred from the user's prior privacy judgments; otherwise, the request is escalated to the user (Figure 1).

We evaluate ARIEL against LLM-based reasoning baselines that rely on either general privacy norms or prior user privacy judgments. These evaluations were conducted using state-of-the-art LLMs—including GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4.5—on publicly available datasets covering smart home assistant [1] and educational [55] data-sharing contexts. We find that (1) grounding privacy judgments in prior user privacy judgments yields significantly better personalization than relying solely on general privacy norms, and (2) ARIEL outperforms pure LLM-based reasoning by reducing the F1 score error of appropriate judgments by up to **40.6%**. Our findings indicate that ARIEL is effective at correctly judging requests where the user would approve data sharing. We summarize **our contributions** below:

(1) We demonstrate the limitations of aligning LLM agents with general privacy norms, as well as the shortcomings of current LLM-based approaches that use prior user judgments for personalized data-sharing decisions.
(2) We propose ARIEL, a framework for grounding the agent's privacy judgments in prior user privacy judgments through logical entailment with LLM-generated ontologies. Jointly leveraging LLMs and logical entailment overcomes the rigidity of traditional rule-based systems without incurring the inconsistency inherent in pure LLM-based reasoning.
(3) We show that ARIEL outperforms LLM-based reasoning for personalized privacy decisions, while incorporating user agency and interpretability. Notably, we find that performance gains are larger with smaller models (Gemma 3 4B) and that ARIEL remains effective with as few as 20 prior requests.

2 Related Work

We review closely-related work on personalizing LLM-based agents to make privacy decisions aligned with user privacy preferences. We discuss (1) prior work in developing personal privacy assistants and (2) evaluations of privacy norms and user preferences in LLMs. Finally, we examine works in (3) neurosymbolic reasoning and their relation to ARIEL.

2.1 Personal Privacy Assistants

Personal Privacy Assistants (PPAs) aim to reduce the user mental burden in making privacy decisions [19, 56]. They explore automating mobile app permissions, predicting user privacy preferences, and making information-sharing decisions [45]. Earlier PPAs targeting mobile app permissions clustered similar user privacy preferences to derive privacy profiles that help determine mobile app permissions for individual users [37, 38]. Follow-up works utilized contextual factors, such as device type, requesting app, and the permission type, as features for predicting privacy preferences using traditional ML techniques like logistic regression or collaborative filtering [4, 10, 46, 65, 66]. Other works have proposed PPAs that manage data-sharing decisions by decomposing scenarios into Contextual Integrity (CI) information flows. These systems then utilize rule-based [73, 74] or ML-based [33] approaches to infer the privacy decision. Finally, Natural Language Inference (NLI)-based models have been proposed to detect mismatches between different privacy policies or between policies and user privacy preferences [3, 28, 53].

While significant progress has been made for PPAs, they remain limited in that they require aggregating all user data for training, which requires users to upload their privacy preferences and prior decisions to a centralized entity. Moreover, these methods struggle with generalizing to different domains.

2.2 Contextual Privacy Awareness in LLMs

Recent work has evaluated LLMs on privacy-decision scenarios derived from general privacy norms, while also evaluating alignment of LLM privacy decisions with user privacy preferences.

General Privacy Norms. Recently, researchers studied whether LLMs can judge the appropriateness of data-sharing scenarios that are based on CI [49, 50]. In particular, Mireshghallah et al. [42] and Shao et al. [54] developed ConfAIde and PrivacyLens, respectively, which are datasets that evaluate whether LLMs can reason about sharing data. Yi et al. [72] studied ambiguity in these datasets to incorporate contextual expansions for improved privacy reasoning. Furthermore, Lan et al. [34] improved LLM evaluation on PrivacyLens through post-training on a synthetic dataset. More recently, Li et al. [35] developed a benchmark targeting legal compliance. And Mireshghallah et al. [43] evaluates CI compliance of LLMs with persistent memory that was synthetically generated. These datasets encompass privacy norms sourced from either legal statutes or aggregated judgments from crowdsourced studies, both of which capture the general public's privacy expectations on whether certain data sharing is appropriate in a particular context. Other works have operationalized the contextual awareness of LLMs to reduce privacy leakage of user tasks. Ghalebikesabi et al. [23] developed strategies for LLM-based assistants to be CI compliant in form filling tasks. Bagdasarian et al. [9] implemented a minimization module with an LLM to generate a minimized set of data based on the user's task. Ngong et al. [48] utilized LLMs to reduce private information within a user's prompt. While evaluating LLMs on general privacy norms is important, these norms fail to capture the nuances of individual user preferences [12, 26, 59, 75]. Consequently, reliance on general norms alone is ineffective for developing agents tasked with making personalized privacy decisions.

User Privacy Preferences. Groschupp et al. [25] and Wu et al. [67] investigated personalized access control in LLMs, where the LLM must determine whether to grant data access by evaluating an incoming request against a user's access-control preferences (e.g., "I prefer not to share data unless necessary"). However, it might not be practical to assume that users can explicitly specify access-control preferences. Further, delegating the access control decision to LLMs may result in misalignment with the user's access-control preferences, especially with limitations of LLMs such as hallucinations [30, 31, 76] and biases [22]. Another work studied how well LLMs align with human preferences in household human-robot interaction contexts [57]. Our work, on the other hand, focuses on a different problem setup, where we evaluate LLMs making privacy decisions on behalf of users. The final work evaluates how well judgments from LLMs align with user expectations [15]. However, their evaluation is quite limited as the dataset is synthetically generated with the authors annotating the user-appropriateness labels.

2.3 Neurosymbolic Approaches for Reasoning

Prior works that are methodologically similar to ours broadly fall under neurosymbolic reasoning, specifically where the LLM acts as an interface to user inputs and offloads the reasoning to rule-based/logical approaches. Such works include using an LLM to translate natural language into a formal intermediate language, which is then offloaded to a deterministic solver to perform the logical reasoning [51, 52, 60, 70]. While ARIEL does not employ a theorem solver, it does use few-shot prompting with an LLM to map data-sharing requests to levels in a corresponding ontology, which are then passed through a rule-base component to determine entailment.

3 Personalizing Privacy Decisions

As illustrated in Figure 1, a personal agent executes tasks on a user's behalf. The agent has access to the user's data and manages data-sharing requests from third parties required to complete these tasks. Below, we formalize the problem setup describing how the agent determines whether transmitting user data is appropriate.

3.1 Problem Statement

We consider an LLM-based agent A_θ that leverages persistent memory D_u [16, 68] to make personalized privacy decisions for a user u. We assume a non-adversarial setting where the data-sharing requests R_u are either authenticated or otherwise not maliciously designed to manipulate the agent A_θ. Upon receiving a request R_u to share data d_u, the agent must infer the user's judgment l_u:

$$A_\theta(D_u, R_u) \in \{\text{appropriate}, \text{inappropriate}\}.$$

The memory D_u represents a history of the user's past privacy judgments l_u^i on data-sharing requests R_u^i:

$$D_u = \{P_u^i\}_{i=1}^n, \quad \text{where } P_u^i = (R_u^i, l_u^i). \tag{1}$$

To illustrate the problem setup, consider the following scenario:

Example 3.1. Let the incoming request R_u be from a bank asking for the user's partial Social Security Number (SSN) to open a checking account. The memory D_u contains a prior request R_u^i in which the same bank asked for the user's full SSN for the same purpose, which the user judged as appropriate (l_u^i = `appropriate`). The agent must decide whether it can infer the appropriateness of R_u based on this precedent (R_u^i, l_u^i).

We represent each data request as a CI information flow [49, 50] containing five parameters: (1) *data type*, the data that is being transmitted; (2) *data subject*, the owner of the data; (3) *data sender*, the individual sharing the data; (4) *data recipient*, the individual receiving the data; (5) *transmission principle*, the purpose and/or condition under which the data is shared. We formally define a request below:

Definition 3.1 (Request)**.** Define $R_u = \{d_u, u, s, r, t\}$ as a set containing five parameters where d_u is the user's data, u is the data subject, s is the data sender, r is the data recipient, and t is the transmission principle.

Using Definition 3.1, we can map the prior and incoming requests from Example 3.1 to the corresponding five parameters, shown in Figure 2.

Prior Request	Incoming Request
data type: full SSN	**data type:** partial SSN
data subject: user	**data subject:** user
data sender: agent	**data sender:** agent
data recipient: bank	**data recipient:** bank
transmission principle: open checking account	**transmission principle:** open checking account
judgment: appropriate	**judgment:**

Figure 2: A example mapping of the prior request with user judgment and the incoming request from Example 3.1 to the five parameters.

3.2 Baseline Approaches

Using an LLM to address the privacy decision problem defined above requires addressing a key challenge: how to design the prompt for the LLM. Here, we discuss two baseline approaches to prompt the LLM to make privacy judgments, along with their limitations.

Privacy Norms-based Baseline. The first choice for prompting the LLM is to utilize Privacy Norms. In this approach, the prompt contains three parts. The first part is the instruction to the LLM to make privacy decisions, the second part includes the privacy norms, and the third part is the incoming request. These privacy norms, akin to CI-inspired approaches, can include the *general population's* overall acceptability of a particular action [20]. For instance, determining whether a user should share a partial SSN with a bank would rely on broad societal acceptability. However, such an approach is limited because general norms lack the granularity required to capture user-specific, personalized privacy decisions.

Privacy Preferences-based Baseline. To address this limitation, an alternative approach is to prompt the LLM explicitly with the user's privacy preferences. In this approach, the prompt also includes three parts: the general task instruction, the user's privacy preferences, and the incoming request. These preferences can be based on rules or policies that the user specifies, similar to Groschupp et al. [25], or they can include a list of the user's privacy judgments on prior requests (e.g., leveraging a past decision about sharing a full SSN to infer the judgment of incoming request) [18, 65].

Limitations. The prompting approaches discussed above, whether based on general norms or prior user privacy decisions, exhibit limitations in terms of (1) alignment, (2) interpretability, and (3) user agency.

Appropriateness judgments generated directly by LLMs remain unreliable due to limited logical reasoning capabilities [36] and the potential for hallucinations that deviate from system prompts [30, 31, 76]. Consequently, these judgments are not guaranteed to *align* with user privacy preferences, resulting in errors that limit effective personalization.

Furthermore, the decision-making process is opaque. Even when reasoning traces are elicited, they are often unfaithful to the LLM's actual logic [11, 71]. This limitation prevents users from accurately *interpreting* the reasoning behind the agent's judgment, thereby eroding the user's trust in the agent.

Finally, because these methods do not align with user preferences and are not inherently interpretable, they hinder *user agency*. While personalized agents should aim to mitigate privacy fatigue, this objective must not come at the cost of compromising user awareness of and control over how decisions are made on their behalf [44, 77].

3.3 Requirements

There is a need to develop agents that make privacy decisions while avoiding the limitations mentioned above. In particular, we posit that such agents must satisfy the following requirements:

(1) **Alignment.** The privacy decisions made by the agent on behalf of the user need to accurately reflect the user's prior privacy decisions. Alignment is crucial for effective personalization of agents.
(2) **Interpretability.** The agent needs to provide reasoning that is transparent and traceable so that the user stays in the loop throughout the privacy-decision process. Interpretability maintains user trust in the system.
(3) **User Agency.** In scenarios where the agent lacks enough information from the user's prior privacy decisions to make a judgment, the agent must determine when it should escalate the request to the user for assistance. Importantly, the agent needs to strike the right trade-off between automated privacy decisions and user control.

4 ARIEL: Privacy Judgments with Entailment

We propose ARIEL, a framework enabling aligned, structured, and interpretable reasoning for personalized privacy judgments. ARIEL operates as a specialized module for privacy decisions within an end-to-end agentic system (see Figure 1). We describe the framework in detail below.

4.1 Overview

4.1.1 Entailment. The core principle of ARIEL is to personalize privacy judgments by determining if a prior judgment made by the user necessitates the same judgment for the incoming request:

$$(R_u^i, l_u^i) \implies (R_u, l_u^i)$$ *Does the judgment l_u^i for the prior request R_u^i logically imply the same judgment for the incoming request R_u?*

This approach addresses the limitations of LLM-based baselines, which do not provide a logical way to infer user judgments from prior decisions. In contrast, entailment treats prior judgments as logical premises that constrain future decisions, offering an interpretable and controllable method for making privacy decisions. The following example highlights how entailment enables privacy decisions based on prior requests.

Example 4.1. Referring back to Example 3.1: a user previously approved sharing their *full* SSN to open a bank account (l_u^i = `appropriate`). The incoming request asks for a *partial* SSN in the exact same context. Logically, the user's approval to share the more sensitive data (full SSN) entails approval to share the less sensitive data (partial SSN).

The main challenge toward applying entailment is the semantic gap between incoming and prior requests, especially if there is no exact lexical overlap between the parameters of the prior and

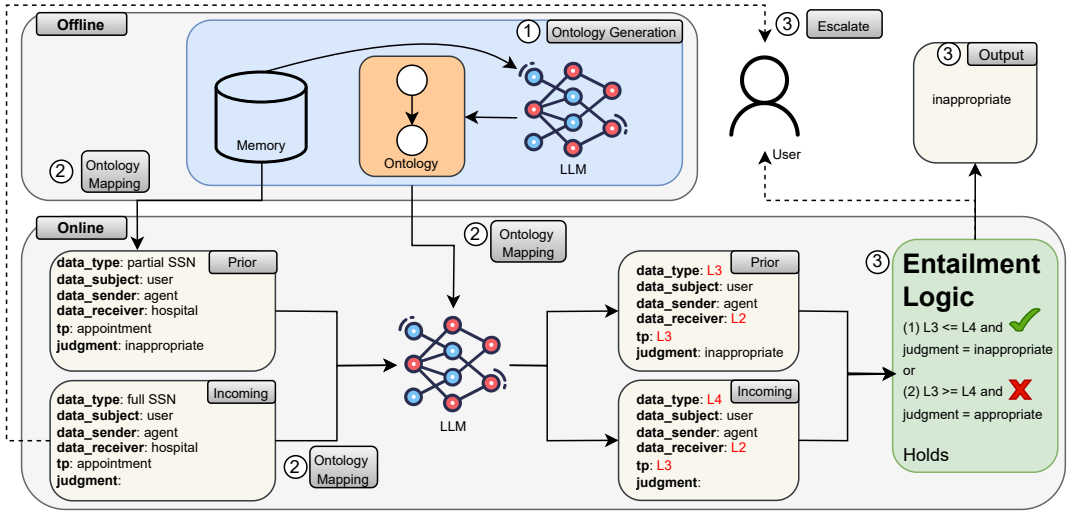

Figure 3: High-level overview of ARIEL, which can be broken down into three components. ① ARIEL uses an LLM to generate an ontology for each user based on the user's prior privacy judgments on data-sharing requests. ② Once ARIEL receives an incoming request, it goes through each prior request from memory to determine if entailment holds. The LLM maps the parameters in both requests to corresponding levels in the generated ontologies. ③ A set of rules is applied to determine whether entailment holds between each mapped prior request and the mapped incoming request. If the incoming request is not entailed by any prior requests, ARIEL escalates the incoming request to the user.

incoming requests. ARIEL addresses this gap using LLM reasoning and rule-based components unified by ontologies. Drawing on prior work in mobile privacy analysis [5, 6, 63], we utilize ontologies to capture entity relationships between parameters in the incoming and prior requests. For example, we can capture the hierarchy of data sensitivity within an ontology where partial SSN is less sensitive than full SSN. However, we simplify the standard graph structure of ontologies into directed paths, where each vertex represents an entity level (detailed in Section 4.2).

4.1.2 Design. The architecture of ARIEL, illustrated in Figure 3, consists of three components:

(1) **Ontology Generation:** use an LLM to generate relevant ontologies for each user over the parameters in each prior request and incoming request.
(2) **Ontology Mapping:** map the parameters of the incoming request and relevant prior requests to levels in the generated ontologies.
(3) **Entailment Logic:** apply a set of entailment rules to determine if the user's prior judgments entail the incoming request.

The three components of ARIEL operate in two phases. In the **offline phase**, *Ontology Generation* creates ontologies for each user based on the user's prior privacy judgments and the LLM's societal norm awareness. The **online phase** processes each incoming request: first, *Ontology Mapping* selects relevant prior requests with user judgments from memory, then maps the parameters in each request to the generated ontologies. This step provides a unified representation of the parameters in both the incoming and prior requests, enabling the entailment step. Next, *Entailment Logic* applies a set of rules to determine if entailment holds between each mapped prior request and the mapped incoming request. If the incoming request is not entailed by any prior requests, ARIEL escalates the incoming request to the user.

4.1.3 Meeting Requirements. ARIEL satisfies the design requirements outlined in Section 3.3 as follows:

(1) **Alignment:** Appropriateness judgments are grounded strictly in the user's prior privacy judgments via logical entailment. The prior requests are mapped onto ontologies, which are generated from the user's prior privacy judgments.
(2) **Interpretability:** All decisions are transparent and auditable. Users can trace ARIEL's judgment back to the specific prior request and the logical rules used to establish the entailment.
(3) **User Agency:** ARIEL maintains user agency by escalating decisions only when an incoming request is not logically entailed by prior judgments. This prevents the agent from making appropriateness judgments when the user's prior privacy judgments are insufficient. Furthermore, by providing interpretable and aligned privacy judgments, ARIEL ensures users retain awareness of and control over the automated decision-making process.

4.2 Ontology Generation

We first describe how ARIEL obtains a set of ontologies O. The **challenge** in generating ontologies is that the agent lacks access to all possible values for each parameter in a request, and it is infeasible to anticipate all these values. For example, a *data recipient* ontology must be sufficiently comprehensive to allow ARIEL to accurately map various recipients within a request to the ontology.

To address this challenge, ARIEL relies on limited initial information: the domain of the incoming request (e.g., an education context) and the user's prior privacy judgments D_u. Leveraging these two sources, the LLM generates user-specific ontologies during an offline phase. The LLM's ability to construct ontologies from limited information enables ARIEL to adapt to new domains without substantial modification.

ARIEL instructs the LLM to generate an ontology for each request parameter, excluding the *data subject*, which is always the same for each request. The LLM generates each ontology based on a hierarchical relationship. We detail each ontology with its corresponding relationship below (please refer to Appendix A.5 for the exact ontology generation prompts):

(1) *Data Type.* Represents data sensitivity as perceived by the user. Sensitivity **increases** at higher levels (Level 1 is least sensitive).
(2) *Data Sender.* Represents the sender's authority to transmit data. Authority **decreases** at higher levels (Level 1 has most authority).
(3) *Data Recipient.* Represents the user's trust in and the authority of the recipient. Trust and authority **decrease** at higher levels (Level 1 has highest trust/authority).
(4) *Transmission Principle.* Represents the transmission purpose and safeguards. Safeguard strength and purpose relevance **decrease** at higher levels (Level 1 has the strongest safeguards/relevance).

Each hierarchical relationship is derived from two sources: (1) **User's prior privacy judgments:** Relationships are inferred directly from the user's prior privacy judgments (D_u). For instance, if a user approves a request involving parents but denies an identical request involving friends, ARIEL infers a higher trust level for parents. This ensures the ontology is personalized. (2) **General norms:** For relationships not present in D_u, ARIEL leverages the LLM's societal norm awareness. For example, the LLM can infer that a typical user generally trusts family more than acquaintances. Figure 4 illustrates a data type ontology generated by Gemini 2.5 Pro for a particular user in the evaluation dataset using this method.

***data type* Ontology**
L1. Non-personal, publicly available information.
L2. User preferences and habits that are not directly identifiable.
L3. Information about the user's home environment and security.
L4. Contact information and social connections.
L5. Sensitive personal and location information.
L6. Highly sensitive health and medical data.

Figure 4: Example of generated *data type* ontology from Gemini 2.5 Pro.

4.3 Ontology Mapping

In the online phase, ARIEL evaluates an incoming request R_u by comparing it against prior requests. The **challenge** is that the privacy contexts can diverge significantly if requests differ by multiple parameters simultaneously, making direct entailment unreliable. For example, sharing full SSN with a bank to open a checking account is an entirely different context than sharing full SSN with an IRS agent to file taxes. Furthermore, even when requests are similar, the agent requires a structured process for comparing differing parameter values based on some hierarchical relationship in order to determine the entailment (e.g., comparing *data recipients* of parents vs friends).

To address these challenges, ARIEL employs a structured approach leveraging the generated ontologies O, as detailed in Algorithm 1:

Neighboring Requests. To ensure that comparisons are made between contextually similar requests, we first select prior requests that 'neighbor' the incoming request, denoting the neighboring set as E. More precisely, E contains prior requests R_u^i whose Hamming distance d from the incoming request R_u is at most one. This is equivalent to ensuring that both requests differ by only one parameter (line 4 in Algorithm 1).

Ontology Mapping. To rigorously compare the differing parameters, we introduce an operator $\leq_O$ to formalize the relationship between the parameters of two distinct requests, R and R', based on their mapped levels within O. Specifically, $R' \leq_O R$ holds if every parameter in R' is ontologically "less than or equal to" the corresponding parameter in R. We formally define this as:

Definition 4.1 (Ontological Relationship)**.** Let $R = \{d_u, u, s, r, t\}$, $R' = \{d'_u, u, s', r', t'\}$ be two requests defined in Definition 3.1 that are partially ordered by the set of ontologies O. Then $R' \leq_O R$ holds if, for each corresponding parameter $p \in R$ and $p' \in R'$, either $p' <_o p$ or $p' =_o p$ under the relevant ontology $o \in O$.

Then ARIEL iterates through each neighboring prior request R'_u. For each pair, the agent retrieves the differing parameter values, p (from R_u) and p' (from R'_u). These values are then mapped to the relevant ontology $o \in O$ (line 10 in Algorithm 1). For this ontological mapping, we use the LLM θ to determine the relevant ontology o and map p and p' to their corresponding levels in o.

4.4 Entailment Logic

Once the requests are mapped to the ontology, ARIEL applies a set of rules to determine the judgment of R_u. This involves comparing the mapped levels between the incoming request and each prior request using $\leq_O$, and considering the user judgment l_u^i on the prior request R_u^i. We formally define our rule-based approach for determining entailment below:

Definition 4.2 (Entailment)**.** Let R, R' be two requests with l' being an appropriateness label for R'. Let O be a set of ontologies. We say that the entailment $(R', l') \implies (R, l')$ holds if one of the following two cases holds:

(1) $l' =$ `inappropriate` and $R' \leq_O R$
(2) $l' =$ `appropriate` and $R \leq_O R'$

ARIEL checks if entailment holds using Definition 4.2 (line 11 in Algorithm 1). We note that multiple neighboring requests can be

Algorithm 1 Online phase of ARIEL

Require: LLM θ, user u, user knowledge base D_u, incoming request R_u, set of ontologies O
Ensure: appropriate, inappropriate, or undetermined

1: $E = \{\}$ ▷ Set of prior request neighboring R_u
2: $c_+ \leftarrow 0, c_- \leftarrow 0$
3: **for** $(R_u^i, l_u^i) \in D_u$ **do**
4: **if** $d(R_u^i, R_u) \leq 1$ **then** ▷ Requests differ by one parameter
5: $E = E \bigcup \{(R_u^i, l_u^i)\}$
6: **end if**
7: **end for**
8: **for** $(R'_u, l'_u) \in E$ **do**
9: Let $p \in R_u$, $p' \in R'_u$ such that $p \neq p'$ ▷ Differing parameter
10: $p_O \leftarrow \theta(O, p)$, $p'_O \leftarrow \theta(O, p')$ ▷ Ontological mapping
11: **if** $l'_u ==$ inappropriate $\wedge\ p'_O \leq p_O$ **then**
12: $c_- \leftarrow c_- + 1$
13: **else if** $l'_u ==$ appropriate $\wedge\ p_O \leq p'_O$ **then**
14: $c_+ \leftarrow c_+ + 1$
15: **end if**
16: **end for**
17: **if** $c_+ > c_-$ **then** ▷ Determine Judgment
18: **Return** appropriate
19: **else if** $c_- > c_+$ **then**
20: **Return** inappropriate
21: **else**
22: **Return** Undetermined ▷ No majority, undetermined
23: **end if**

entailed, with potentially opposing judgments. This could be due to inconsistency in the user's prior judgments or incomplete information from the ontology. To address this, the agent goes through each neighboring request and keeps a running vote of the appropriate and inappropriate judgments that were entailed. Whichever appropriateness label receives the majority vote is the final judgment. If there is no majority, the agent outputs 'undetermined' (line 17 in Algorithm 1), signifying that the incoming request should be escalated to the user for further assistance.

Lastly, we note that in scenarios where multiple appropriate and inappropriate prior requests are entailed, ARIEL can be modified so that the agent to escalate the incoming request to the user instead of employing a majority vote.

5 Evaluation Setup

We present the experimental framework designed to assess the efficacy of ARIEL. In the following sections, we detail the specific datasets used, describe the baselines compared, and define the evaluation metrics employed to quantify performance.

5.1 Datasets for Data-sharing Requests

We use existing datasets that elicit privacy preferences through factorial vignette surveys based on CI. Participants are presented with data-sharing scenarios and asked to indicate their acceptance of each one [1, 7, 8, 13, 29, 40, 55]. To simulate a user with prior privacy judgments, each data-sharing scenario is converted into a request (Definition 3.1) with appropriateness judgments directly received from the participants. Then, we divided a participant's set of answered requests into two pairwise disjoint subsets. The first subset was treated as observed prior requests with the user's answers as the judgments. The second subset represents incoming requests where the participant's answers were held out as the ground truth for evaluation. Our evaluation employs two such datasets, which we repurposed for our setup as detailed below. See Figure 5 for example requests from both datasets.

5.1.1 Smart Home Personal Assistant (SPA) dataset. The SPA dataset[1] [1] consists of factorial vignette surveys covering different data-sharing scenarios to understand privacy norms in the SPA ecosystem. The scenarios typically involve a voice assistant provider sending user data to individuals/entities who have access to the user's SPA ecosystem, under a certain purpose and condition. The data subject of a request is always "user" and the data sender is always "assistant provider". The total number of participants is 1,730, with each participant randomly assigned to answer a subset of the questions. This resulted in approximately 180 different scenarios answered by each participant.

Due to the relatively large number of participant responses from the SPA dataset, we restrict our evaluation to a set of randomly sampled 500 users. To ensure a consistent experimental setup across all users, we standardized the response size by randomly selecting 70 responses per user. For each of the 500 users, we partitioned the responses into 60 prior requests and 10 incoming requests. This yields a total of $500 * 10 = 5000$ incoming requests, consisting of 1,765 appropriate and 3,235 inappropriate labels. We also perform an ablation study on varying the number of prior requests in Section 6.2.

Lastly, in the dataset, the participant responses are on a 1-5 Likert scale where 1 and 2 correspond to strongly and somewhat unacceptable responses, respectively, 3 corresponds to neutral, and 4 and 5 correspond to somewhat and strongly acceptable responses, respectively. We convert the responses with 1-2 scores to "Inappropriate" and the responses with 4-5 to "Appropriate". We filter the remaining responses with scores of 3.

5.1.2 Education Dataset. The Education dataset[2] [55] consists of factorial vignette surveys involving the transmission of student data within an educational context. For example, a student's grades are shared by a professor to parents if the student is performing poorly. The data subject of a request is always 'student'. In total, there were 451 respondents, each answering a random subset of the questions which was around 88 and 89.

The participant responses in the dataset contain numerical values, where 1 corresponds to Yes, 2 corresponds to No, and 3-5 correspond to Does not Make Sense options. Specifically, 3 is "The sender is unlikely to have the information", 4 is "The receiver would already have the information", and 5 is "the question is ambiguous." We convert responses with 1 to "Appropriate" and with 2 to "Inappropriate". We filter out the remaining responses with 3-5 scores. Lastly, we filter out users who have fewer than 70 appropriateness judgments, so each user has 60 prior requests and 10 incoming

[1]Dataset can be accessed here: https://osf.io/63wsm/overview
[2]Dataset can be accessed here: https://yansh.github.io/papers/HCOMP/

Dataset	# Sampled Users	# Requests per User	Total	# App	# Inapp
SPA	500	10 incoming 60 prior	5000	1765	3235
Education	302	10 incoming 60 prior	3020	1378	1642

Table 1: A summary of the statistics of the evaluation datasets after our pre-processing. This contains the sampled number of users (# Users), sampled number of requests per user (# Requests per User), the total number of incoming requests among all users (Total), the number of incoming requests with appropriate labels (# App), and the number of incoming requests with inappropriate labels (# Inapp).

requests. In total, there are 302 users and $302 * 10 = 3020$ total incoming requests, consisting of 1378 appropriate labels and 1642 inappropriate labels.

We present a summary of the statistics of the evaluation datasets after our pre-processing in Table 1.

SPA Dataset
data type: email content
data subject: user
data sender: assistant provider
data recipient: partner
transmission principle: reading emails on user's behalf

Education Dataset
data type: grades
data subject: student
data sender: professor
data recipient: parents
transmission principle: if student is performing poorly

Figure 5: Example requests from the SPA and Education Dataset that are provided to the LLMs in our evaluations.

5.2 Baselines

Following the description in Section 3.2, we employ two natural baseline prompt strategies: (1) privacy norms representative of the general population and (2) prior privacy judgments from users.

5.2.1 Privacy Norms-based Baseline. Privacy norms capture the general population's accepted appropriateness of information flows. We consider two prompts.

(1) *zero-shot*. We construct the prompt to utilize the out-of-the-box privacy norm awareness of LLMs, following from recent prior work on evaluating CI-compliance in LLMs [42, 54]. The prompt directs the LLM to judge the appropriateness of each information flow based on its knowledge of general privacy norms. The prompts can be found in Appendix A.1.
(2) *privacy norms*. We construct the prompt such that the LLM's privacy judgment is now based on a provided subset of representative privacy norms. These extracted privacy norms are taken directly from the analysis provided in the original papers for each dataset. Table 2 contains a subset of the general privacy norms from both evaluation datasets that were used by the LLMs. The prompt directs the LLM to judge the appropriateness of each information flow based on the provided general privacy norms. The prompts containing the entire set of privacy norms can be found in Appendix A.2.

Rule	Judgment
SPA Dataset, user recipients	
data type: voice command history	Inappropriate
data type: bank account details	Inappropriate
recipient: neighbors	Inappropriate
recipient: visitors in general	inappropriate
SPA Dataset, non-user recipients	
data type: bank account details	Inappropriate
recipient: advertising agency	Inappropriate
tp: no purpose/condition	Inappropriate
Education Dataset	
sender: professor; **recipient**: graduate schools; **data type**: grades; **tp**: with subject's consent	Appropriate
sender: TA; **recipient**: classmates; **data type**: grades; **tp**: subject performing poorly	Inappropriate

Table 2: A subset of the top privacy norms for user and non-user recipients from the SPA dataset, and the top appropriate and inappropriate norms from the Education dataset. The LLMs are provided either the general norms from the SPA or education dataset, depending on which dataset is being evaluated. We denote tp as transmission principle.

5.2.2 Privacy Preferences-based Baseline. We provide privacy judgments on prior requests made by a user. We consider two prompts.

(1) *In-Context Learning (ICL)*. We construct the prompt to leverage ICL [14] by instructing the LLM to infer the user's privacy preferences from their judgments on prior requests D_u. When the LLM judges an incoming request R_u, its decision is therefore personalized based on the inferred preferences in D_u. These prompts are in Appendix A.3.
(2) *ICL (w Undet)*. We modify the instruction in *ICL* to output 'undetermined' if the model is unable to confidently judge whether the incoming request is appropriate or inappropriate based on the user's privacy preferences. This modification accounts for scenarios where the user's privacy preferences may not always provide sufficient information to accurately judge the incoming request. The prompts are in A.4.

5.3 Models and Evaluation metrics

Models. We evaluate ARIEL and the baselines on three advanced models: Gemini 2.5-Pro (gemini-2.5-pro), GPT-5 (gpt-5-2025-08-07), and Claude Sonnet 4.5 (claude-sonnet-4-5-20250929). Additionally, we include a smaller, open-source model, Gemma 3 4B IT [58].

Metrics. We compare the agent's predicted judgments against each sampled user's judgments on incoming requests. We report separate F1 scores for both the *appropriate* and *inappropriate* classes. Given an evaluation dataset D containing incoming requests R_u

with judgments l_u for each user u, we define the precision P_A for the appropriate class (c^+ = `appropriate`) as:

$$P_A := \frac{|(R_u, l_u) \in D : A_\theta(D_u, R_u) = c^+ \wedge l_u = c^+|}{|(R_u, l_u) \in D : A_\theta(D_u, R_u) = c^+|}.$$

We also define recall R_A for the appropriate class as:

$$R_A := \frac{|(R_u, l_u) \in D : A_\theta(D_u, R_u) = c^+ \wedge l_u = c^+|}{|(R_u, l_u) \in D : l_u = c^+|}.$$

Then we can define the F1 score $F1_A$ for the appropriate class as:

$$F1_A = 2 \cdot \frac{P_A \cdot R_A}{P_A + R_A}.$$

A similar process can be used to obtain the F1 score for the inappropriate class $F1_I$. High $F1_A$ indicates the agent is correctly sharing user data (utility), whereas high $F1_I$ indicates the agent is correctly withholding data (privacy).

Evaluation Procedure for general norms. For each unique request in the datasets, we generate a prompt as described in Section 5.2.1. This results in 825 prompts for the SPA dataset and 1411 for the Education dataset, covering both *zero-shot* and *privacy norms*. Once the prompts are generated, we provide all of them to the models to obtain their responses. Then, for each user, we compare their judgment on an incoming request with the LLM's judgment on that same request across both general norms prompt types.

Evaluation Procedure for user privacy preferences. We generate a prompt for each incoming request with the user's privacy judgments on prior requests. This results in 5000 prompts for the SPA dataset and 3020 prompts for the Education dataset, covering *ICL*, *ICL (w Undet)*, and `ARIEL`. Incoming requests resulting in an 'undetermined' judgment are excluded from the evaluation for the respective baselines (i.e., *ICL (w Undet)* or `ARIEL`). To account for this partial coverage, we report the Support (Sup) metric, defined as the total number of appropriateness judgments (i.e., "appropriate" or "inappropriate") made by the model.

6 Results

We present experiments, ablation studies, and error analysis to study the efficacy and robustness of `ARIEL`. First, we evaluate the end-to-end performance improvement of `ARIEL` over existing baselines in Section 6.1. Then, we evaluate how varying the number of prior user judgments affects `ARIEL` in Section 6.2. Finally, we qualitatively analyze the failure modes of `ARIEL` in Section 6.3.

6.1 End-to-End Evaluation

We show the results for all approaches in Table 3. Based on these results, we make the following observations:

(1) `ARIEL` *enhances personalized privacy judgments compared to pure LLM reasoning.* `ARIEL` consistently outperforms LLM-based baselines across all datasets and reasoning models, with notable gains in appropriate judgments. For example, `ARIEL` improves upon *ICL (w Undet)* by **10.4%** $F1_\text{A}$ with GPT-5 on the Education dataset. In other words, the $F1_\text{A}$ error reduces by **40.6%**, demonstrating that logical entailment effectively grounds judgments in prior user privacy judgments. Notably, these improvements are magnified in smaller models. With Gemma 3 4B IT, `ARIEL` yields $F1_\text{A}$ gains of 13% and 17% over *ICL (w Undet)* on the SPA and Education datasets, respectively. This demonstrates `ARIEL` is suitable for on-device deployment of personalized agents.

(2) `ARIEL` *accurately identifies when the appropriateness of an incoming request is challenging to judge.* We observe that the number of judgments that were undetermined by `ARIEL` is higher than *ICL (w Undet)*. For Gemini 2.5 Pro, we observe that `ARIEL` judged 1203 and 1339 incoming requests as undetermined for SPA and Education datasets, respectively, compared to just 387 and 695 for *ICL (w Undet)*. This observation raises the question of whether the appropriateness label of the incoming requests left undetermined by `ARIEL` can actually be inferred by an LLM based on the user's prior privacy judgments.

To investigate this question, we analyzed the *ICL* performance on two subsets of incoming requests: those where `ARIEL` found logical entailment, and those where it did not (see Table 4). We observe that ICL achieves at least a 10% increase in both $F1_\text{A}$ and $F1_\text{I}$ on the entailed requests compared to the non-entailed ones. As such, `ARIEL` accurately identifies requests that are answerable based on prior judgments, effectively escalating only the more challenging ones. Conversely, *ICL (w Undet)* produces fewer 'undetermined' outputs, suggesting that LLMs struggle to recognize when they lack sufficient information to make a privacy judgment. This observation aligns with findings in other domains regarding LLMs' inability to abstain when lacking necessary knowledge [21, 32].

(3) *Relying only on the LLM's awareness of privacy norms is insufficient for personalization.* We observe that performance on appropriate judgments from *zero-shot* is low for both models and datasets. For inappropriate judgments, *zero-shot* prompting can perform much better than on appropriate judgments, but the results remain suboptimal. Hence, the LLM's awareness of privacy norms cannot sufficiently capture the nuanced behavior of user privacy preferences within the SPA and Education contexts.

(4) *General privacy norms extracted from overall user responses for each dataset can improve personalized privacy judgments.* When comparing *privacy norms* and *zero-shot*, we observe that providing privacy norms to the LLMs can improve appropriate judgments for the SPA dataset across all models, and improve inappropriate judgments for the Education dataset on most models. However, including general norms can sometimes degrade performance. For instance, with Gemini 2.5 Pro on the Education dataset, $F1_\text{A}$ decreases by 22% while $F1_\text{I}$ improves by only 10%, suggesting that adding general privacy norms is not always beneficial.

(5) *Prior user privacy judgments improve personalized privacy judgments over general privacy norms.* By comparing the $F1_\text{A}$ and $F1_\text{I}$ between *ICL* and both *zero-shot* and *privacy norms* for all datasets and models, it is clear that personalizing LLMs with prior user privacy judgments improves appropriate and inappropriate judgments over general privacy norms. In some cases, the improvement over general privacy norms can be as large as **50%** for $F1_\text{A}$.

(6) *Personalized privacy judgments improve when including undetermined as an option.* When comparing *ICL* and *ICL (w Undet)*, we observe an improvement in $F1_\text{A}$ and $F1_\text{I}$ for all LLMs. These

Dataset	Model	zero-shot		privacy norms		ICL			ICL (w Undet)			ARIEL		
		$F1_A$ (%)	$F1_I$ (%)	$F1_A$ (%)	$F1_I$ (%)	$F1_A$ (%)	$F1_I$ (%)	Sup	$F1_A$ (%)	$F1_I$ (%)	Sup	$F1_A$ (%)	$F1_I$ (%)	Sup
SPA	Gemini 2.5 Pro	49.4	75.8	59.1	71.5	81.8	91.0	5000	82.4	92.2	4613	**87.2**	**94.2**	3797
	GPT-5	27.1	78.3	58.4	71.1	79.0	90.6	5000	80.0	91.1	4921	**87.4**	**94.2**	3810
	Claude Sonnet 4.5	40.0	75.0	57.7	73.6	81.4	91.0	5000	83.3	91.7	4900	**86.6**	**93.8**	3824
	Gemma 3 4B IT	45.1	72.8	46.4	73.1	64.1	84.4	5000	62.5	84.2	4999	**75.5**	**88.4**	3680
Education	Gemini 2.5 Pro	60.1	61.7	38.5	71.6	69.9	80.8	3020	75.5	87.2	2325	**84.1**	**88.2**	1681
	GPT-5	62.9	55.6	51.2	69.2	74.0	82.9	3020	74.4	83.2	2961	**84.8**	**89.2**	1603
	Claude Sonnet 4.5	61.4	63.0	62.3	62.5	75.4	82.8	3020	75.0	82.8	2950	**84.0**	**88.4**	1608
	Gemma 3 4B IT	64.1	18.0	46.7	66.8	61.7	78.7	3020	63.4	78.8	3020	**80.4**	**86.1**	1644

Table 3: Evaluating different methodologies involving LLMs on requests from SPA and Education datasets. We report the F1-score for appropriate ($F1_A$) and inappropriate ($F1_I$) class, and the total number of appropriateness judgments (Sup) for *ICL*, *ICL (w Undet)*, and ARIEL. We boldface the best $F1_A$ and $F1_I$ scores, respectively, across all methods for each model and dataset. The evaluations show that ARIEL outperforms all other methodologies for all models on both datasets.

Model	Not Entailed			Entailed		
	$F1_A$ (%)	$F1_I$ (%)	Sup	$F1_A$ (%)	$F1_I$ (%)	Sup
SPA Dataset						
Gemini 2.5 Pro	69.6	79.4	1203	87.0	94.2	3797
GPT-5	60.4	79.2	1190	85.7	94.0	3810
Claude Sonnet 4.5	68.1	79.8	1176	86.5	94.0	3824
Education						
Gemini 2.5 Pro	58.6	74.2	1339	78.7	86.1	1681
GPT-5	65.3	76.4	1417	82.0	88.6	1603
Claude Sonnet 4.5	68.3	76.9	1412	81.8	87.9	1608

Table 4: Analyzing the judgments from *ICL* on two sets of incoming requests: one where none of the prior requests were entailed by ARIEL (Not entailed), and the other where entailment occurred by ARIEL (Entailed). These results suggest that ARIEL accurately identifies when the appropriateness of incoming requests are difficult to judge based on the user's prior privacy judgments.

results suggest that giving an LLM the option to abstain reduces the likelihood that the model forces an erroneous appropriateness judgment.

6.2 Ablation Evaluation

We evaluate how the number of prior requests affects the performance of ARIEL and baseline approaches. Moreover, we investigate how the choice of reasoning affects LLM privacy judgments.

6.2.1 Number of prior requests. We explore how varying the number of prior requests with user judgments affects the agent's judgments. Specifically, we compare the $F1_A$, $F1_I$, and Support (Sup) of ARIEL against *ICL (w Undet)* on the SPA dataset using Gemini 2.5 Pro. As evident in Figure 6, while the $F1_I$ gap remains relatively constant, the disparity in $F1_A$ widens significantly as the number of prior requests decreases. Notably, when reducing the prior requests from 60 to 20, $F1_A$ drops by only 3.4% for ARIEL, compared to an 8.8% decline for *ICL (w Undet)*. This demonstrates that ARIEL is more robust to limited prior privacy judgments. Nonetheless, we observe that the number of entailments generated by ARIEL decreases substantially as the number of prior requests decreases. This result is expected, as there are fewer candidate requests available to satisfy the entailment criteria (i.e., fewer prior requests differ by exactly one parameter from the incoming request).

6.2.2 Assessing the impact of reasoning with Chain of Thought (CoT). To assess the impact of reasoning strategy choice on appropriateness judgments, we implement CoT by prompting the LLM to provide a reasoning before making a judgment [64, 69]. Inspired by findings that CoT affects privacy leakage [42], we evaluate its effect on judgment accuracy across *zero-shot*, *privacy norms*, and *ICL (w Undet)*. Figure 7 indicates that CoT typically provides negligible changes in performance. However, we observe for *zero-shot* with Gemini 2.5 Pro that the $F1_A$(%) and $F1_I$(%) are significantly impacted by CoT, both positively and negatively.

While CoT provides mixed results for general privacy norms, the results clearly demonstrate that adding CoT to both *ICL* and *ICL (w Undet)* consistently hurts their performance. When manually inspecting the generated reasoning traces, we notice that the LLM makes misaligned assertions about the user's privacy judgments, which propagate all the way to the model's final judgment (see Section 6.3 for more information). This result also demonstrates that eliciting explanations from LLM-based reasoning can degrade appropriateness judgments. In contrast, ARIEL provides a clear, traceable reasoning while maintaining well-grounded personalized privacy judgments.

6.3 Qualitative Analysis

Finally, we qualitatively analyze incorrect judgments made by both *ICL (w Undet) + CoT* and ARIEL. We sample 20 examples from each method on the SPA dataset with Gemini 2.5 Pro. The results are detailed in Table 5.

6.3.1 Error Analysis on Appropriateness Judgments for ICL (w Undet) + CoT. Analysis of the reasoning traces reveals that Gemini 2.5 Pro occasionally performs implicit entailment when identifying prior requests that share identical parameters except for one. This behavior is noteworthy, as we did not explicitly instruct the model to use this logic, but only to extract patterns from user privacy judgments.

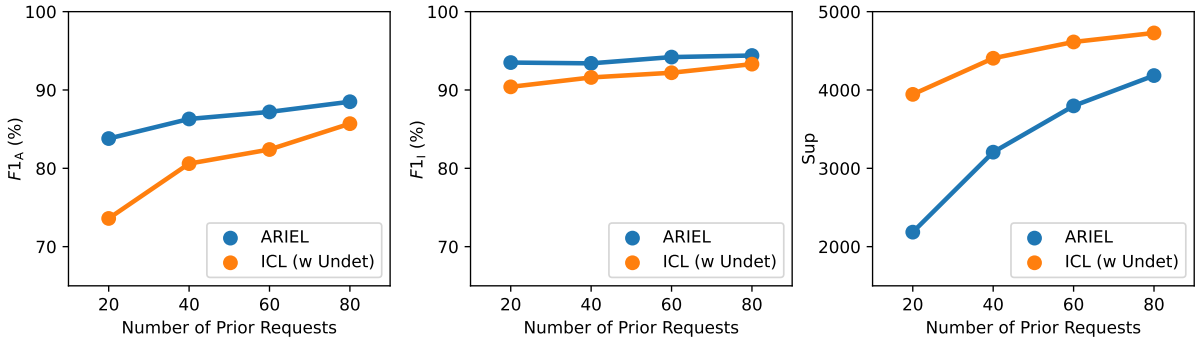

Figure 6: Ablation study on the number of prior requests with user judgments contained in user memory D_u. We evaluate *ICL (w Undet)* and ARIEL on the SPA dataset with Gemini 2.5 Pro. We report the F1-score for appropriate $F1_A$ and inappropriate $F1_I$ class, and the total number of appropriateness judgments (Sup). We find that ARIEL is more robust to varying the number of prior requests compared to *ICL w (Undet)*.

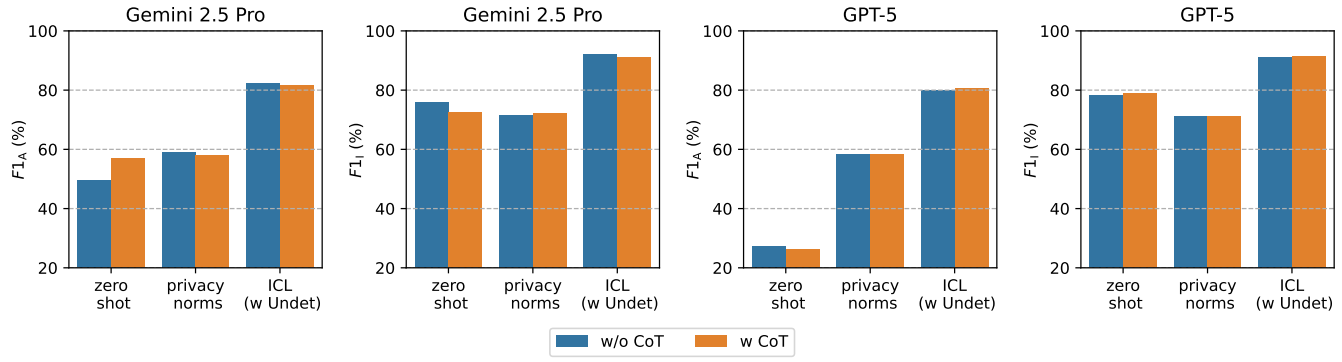

Figure 7: Ablation study on CoT. We evaluate *zero-shot*, *privacy norms*, and *ICL w (Undet)* on the SPA dataset with Gemini 2.5 Pro. We report the F1-score for the appropriate $F1_A$ and inappropriate $F1_I$ class. The results suggest that generally CoT has negligible impact on privacy judgments across both models.

We find that the majority of errors stem from *preference misalignment* (14 examples), where the model draws incorrect equivalencies between distinct parameter values to perform this entailment. For instance, given prior requests involving "close friends" or "parents" as recipients and an incoming request involving "housemates," the model infers that the privacy judgment should remain consistent due to perceived similarity in recipient type. However, the user may not view these three recipients as similar.

We also attribute a subset of errors to the model's inherent conservatism regarding information sharing, which we describe as *default to inappropriate* (5 examples). This occurs when the model concludes it cannot use the most similar prior requests that are judged as "appropriate" by the user. However, rather than outputting "undetermined", the model reasons that the judgment is inappropriate. Finally, we identified one instance of *contradictory prior requests*, where the provided user history contained judgments that logically conflicted with the ground truth for the incoming request.

6.3.2 Error Analysis on Appropriateness Judgments for ARIEL. We classify errors in ARIEL into three categories: (1) ontology generation, (2) ontology mapping, and (3) contradictory prior requests. The majority of errors stem from ontology generation, where the generated ontology produces incorrect entailments (10 examples). This typically arises from data sparsity when specific parameter values (e.g., 'housemates' or 'close friends') appear only once or twice in the user's prior requests. The model struggles to accurately reflect the user's privacy preferences onto the ontology. A second source of error is ontology mapping, where the model fails to map parameter values to the ontology correctly (5 examples). This is often caused by domain-specific idiosyncrasies in the evaluation dataset. For instance, the transmission principle has a specific format combining both purpose and condition, which can be challenging for the model to map. Finally, the remaining errors (5 examples) result from contradictory prior requests. For instance, there was a prior request where the user had judged sharing bank account details with business and finance skills for the purpose of making transactions as inappropriate. The incoming request is identical to the prior, except there is no purpose. Thus, the incoming request

Type	Example/Description	Total
	ICL (w Undet) + CoT	
Preference misalignment	The user has previously approved sharing contacts with close friends and parents for the purpose of looking up contact details. Since housemates are a similar type of recipient, this request is likely to be considered appropriate	14
Default to inappropriate	The user has consistently approved sharing email content with advertising agencies for personalized ads, as long as there are some safeguards in place. Since this request has no safeguards, it is inappropriate.	5
Contradictory Prior Requests	The user has previously judged sharing bank account details with business and finance skills for making transactions as inappropriate without any safeguards. Since the current request lacks a transmission principle, it is likely to be considered inappropriate.	1
	ARIEL	
Ontology Generation	An error caused by the hierarchy of the ontology being incorrect. This is due to incomplete information from the user's privacy preferences. For example, "housemate" only appears once in the prior request which is not enough knowledge to determine how trustworthy housemate is for the user.	10
Ontology Mapping	This typically caused by domain specific issues about the dataset. For SPA, the transmission principle has many combinations for puprose and condition, which can make it hard for the model to map to the correct level.	5
Contradictory Prior Requests	The judgments of the prior requests contradict the judgment of the incoming request.	5

Table 5: The different error types of 20 judgments from *ICL (w Undet) + CoT* and ARIEL on SPA dataset with Gemini 2.5 Pro.

should also be inappropriate, since removing the purpose likely makes the request more inappropriate. Both *ICL (w Undet) + CoT* and ARIEL judged this request inappropriate; however, the user judged the incoming request to be appropriate.

Potential solutions for (1) involve providing more user judgments on prior requests during ontology generation, while solutions for (2) involve providing more information on the domain-specific format of each parameter in the ontology mapping prompt. The errors in (3) represent an open challenge that is discussed in more detail in Section 7.1.

7 Discussion

Based on our results and takeaways, we discuss in more detail the broad implications of this work as well as its limitations.

7.1 Personalization Challenges

In this paper, we consider the user's judgments on prior data-sharing requests as proxies for the user's privacy preferences [18, 65]. Personalizing privacy decisions based on these preferences introduces several challenges that must be considered.

Contradictory Privacy Preferences. Highlighted by the error analysis in Section 6.3, contradictions can arise within the user's own history of prior judgments. For example, the user says it is inappropriate to share their banking information with a third party if their data was confidential, but deems it appropriate if there is no safeguard on the data. In such scenarios, the agent may execute the entailment logic correctly but still reach an incorrect privacy judgment. Consequently, the agent must (1) detect these user privacy judgment contradictions and (2) resolve them, typically by escalating the request to the user. The fundamental difficulty lies in distinguishing whether an incorrect privacy judgment stems from an error in the entailment process of ARIEL or an inconsistency in the user's prior privacy judgments.

Evolving Privacy Preferences. The second challenge involves the dynamic nature of privacy preferences as they can change over time. For example, a user can become more conservative about sharing their data as they get older. Currently, ARIEL does not support automatic detection of privacy preference changes. Hence, future adaptations of ARIEL require a mechanism to detect when the user's privacy preferences are shifting and determine which preferences must be updated. One potential mechanism is to periodically escalate incoming requests to the user to confirm the agent's privacy decision.

7.2 User Agency in Personal AI Agents

While personalized agents reduce the burden of privacy management, they must not remove user agency [44, 77]. This observation motivated the entailment logic of ARIEL, where the agent automates clear decisions that can be entailed but escalates requests where the judgment cannot be easily inferred by the user's prior judgments. Our results confirm that identifying these non-entailed incoming requests drives ARIEL's judgment accuracy gains. Furthermore, by providing transparent reasoning traces for user auditing, ARIEL successfully balances automation with necessary human oversight.

Although ARIEL was designed with user agency as a core requirement, further work is needed to strengthen the interaction between the user and the agent. One promising direction is the development of a dedicated user interface (UI) that enables users to: (1) visualize the reasoning steps behind ARIEL's decisions (e.g., which requests were entailed and how they were mapped in the ontology), and (2) intervene in the decision-making process when necessary. Subsequently, this UI can facilitate user studies to evaluate how well ARIEL's privacy reasoning and judgments align with the user's own privacy reasoning [47]. Furthermore, this UI could be used to evaluate potential solutions for the challenges outlined in Section 7.1, such as user verification of ARIEL's reasoning and decisions to identify contradictions or changes in the user's privacy preferences.

7.3 Personalization and Privacy Norms

Our results in Section 6 suggest that effective personalization of LLM agents requires grounding judgments in prior user privacy judgments, as general norms are insufficient. However, this finding

does not invalidate alignment of LLMs with general contextual norms, as this alignment remains crucial for many data-sharing scenarios [42, 54]. Our work highlights that when personalization is a crucial component for making privacy decisions, then the LLM's privacy judgments should be grounded in user-specific preferences rather than general privacy norms.

7.4 Limitations

We identify four main limitations involving our work.

Ontology Mapping and Generation. ARIEL generates generalizable ontologies for individual fields (e.g., data type) based on the user's prior privacy judgments. However, in practice, the ontology of one field may depend on the values of others. For instance, data sensitivity can vary based on the recipient: a user may consider their income more sensitive than SSN when sharing with a parent, but the sensitivity is reversed when sharing with a friend. Our results in Section 6.1 demonstrate that even without explicitly modeling these conditional dependencies, ARIEL accurately infers privacy judgments. We leave the incorporation of context-conditional hierarchies for future work.

Evaluation Datasets. Our evaluation datasets simulate prior requests by partitioning the user's judgments on requests into two pairwise disjoint sets, where one set stores already observed requests while the other contains new requests. While we believe this reasonably approximates our problem setup, it does not incorporate user decision-making over time. Unfortunately, to our knowledge, there are no existing publicly-available datasets that incorporate user privacy preferences over time. An interesting direction for future work could elicit user privacy preferences over a period of time, rather than giving a one-time questionnaire to participants.

Although we believe our evaluation datasets accurately reflect our problem setup, they were not originally framed as AI agents acting on the users' behalf. This distinction is notable, as prior work indicates that privacy preferences can shift when information-sharing decisions are delegated to a personal AI agent [26]. While recent studies have begun eliciting user privacy preferences regarding AI agents and chatbots [26, 39, 59, 67, 77, 78], these datasets are either (1) not publicly available, or (2) restricted to narrowly scoped scenarios. Consequently, this emerging field would benefit significantly from future work that appropriately releases open datasets on user privacy preferences in agentic data-sharing contexts.

Cold-start Cases. Our evaluation does not consider cold-start cases where the user has not made prior judgments on requests–essentially, no prior privacy preferences. Handling these cold-start cases is an important consideration when designing personalized agents. While falling back to general privacy norms is a potential solution, our results demonstrate they are often insufficient for personalized privacy decisions. An alternative solution could be to develop the minimal set of questions to elicit the user's privacy preferences [27] while offsetting user burden. We leave evaluation on cold-start cases as future work.

Non-adversarial Setting. Lastly, our work assumes a non-adversarial setting which could limit the applicability of ARIEL in open environments. However, ARIEL is designed as a specialized module focused on agent privacy decisions, functioning within a larger system that authenticates inputs. An example is Google Play Store where app developers are verified and users provide feedback on problematic apps, which prevents requests to the agent from being maliciously designed. Thus, defense against adversarial attacks, such as prompt injection [24], is delegated to separate system components that focus specifically on these attacks [41, 61]. Additionally, ARIEL operates on limited structured input containing just five fields, inherently reducing the risk of adversarial attacks.

8 Conclusion

In this work, we explored personalization of LLM-based agents for making privacy judgments. We identified that general privacy norms are insufficient for effective personalization, and LLM-based reasoning with ICL does not guarantee that judgments are faithfully grounded in the user's prior privacy judgments. We addressed these limitations by designing ARIEL to ensure judgments are grounded in the user's prior judgments with clear and traceable reasoning that the user can reference. Our evaluations demonstrate that ARIEL can reduce the F1 score error for appropriate judgments by up to **40.6%** over LLM reasoning. Overall, our findings suggest that jointly leveraging LLMs with logical entailment is a promising direction for designing agentic frameworks for personalized privacy judgments.

Acknowledgments

We would like to thank Adrià Gascón, Sarah Meiklejohn, Eugene Bagdasarian, Daniel Ramage, Zhichen Zhao, and Ku Kogan for their constructive discussions and support. The authors used generative AI-based tools to revise the text, improve flow and correct any typos, grammatical errors, and awkward phrasing. James is supported by NSF award number DGE-1842487. Murali and James are supported by REAL@USC-Meta center.

References

[1] Noura Abdi, Xiao Zhan, Kopo M Ramokapane, and Jose Such. 2021. Privacy norms for smart home personal assistants. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[2] Alessandro Acquisti, Allan Friedman, and Rahul Telang. 2006. Is there a cost to privacy breaches? An event study. *ICIS 2006 proceedings* (2006), 94.

[3] Abdullah R Alshamsan and Shafique A Chaudhry. 2022. Detecting privacy policies violations using natural language inference (nli). In *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, 1–6.

[4] Marc Serramia Amoros, William Seymour, Natalia Criado, and Michael Luck. 2023. Predicting privacy preferences for smart devices as norms. In *The 22nd International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

[5] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. {PolicyLint}: investigating internal privacy policy contradictions on google play. In *28th USENIX security symposium (USENIX security 19)*. 585–602.

[6] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions speak louder than words:{Entity-Sensitive} privacy policy and data flow analysis with {PoliCheck}. In *29th USENIX Security Symposium (USENIX Security 20)*. 985–1002.

[7] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. 2018. Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 2 (2018), 1–23.

[8] Noah Apthorpe, Sarah Varghese, and Nick Feamster. 2019. Evaluating the Contextual Integrity of Privacy Regulation: Parents'{IoT} Toy Privacy Norms Versus {COPPA}. In *28th USENIX security symposium (USENIX security 19)*. 123–140.

[9] Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. 2024. Airgapagent: Protecting

privacy-conscious conversational agents. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 3868–3882.
[10] Natã M Barbosa, Joon S Park, Yaxing Yao, and Yang Wang. 2019. "What if?" Predicting individual users' smart home privacy preferences and their changes. *Proceedings on Privacy Enhancing Technologies* (2019).
[11] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv* (2025), v1.
[12] Louise Barkhuus. 2012. The mismeasurement of privacy: using contextual integrity to reconsider privacy in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 367–376.
[13] Jaspreet Bhatia and Travis D Breaux. 2018. Empirical measurement of perceived privacy risk. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 6 (2018), 1–47.
[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[15] Zhao Cheng, Diane Wan, Matthew Abueg, Sahra Ghalebikesabi, Ren Yi, Eugene Bagdasarian, Borja Balle, Stefan Mellem, and Shawn O'Banion. 2024. Ci-bench: Benchmarking contextual integrity of ai assistants on synthetic data. *arXiv preprint arXiv:2409.13903* (2024).
[16] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413* (2025).
[17] Hanbyul Choi, Jonghwa Park, and Yoonhyuk Jung. 2018. The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior* 81 (2018), 42–51.
[18] Jessica Colnago, Lorrie Faith Cranor, Alessandro Acquisti, and Kate Hazel Stanton. 2022. Is it a concern or a preference? An investigation into the ability of privacy scales to capture and distinguish granular privacy constructs. In *Eighteenth symposium on usable privacy and security (SOUPS 2022)*. 331–346.
[19] Jessica Colnago, Yuanyuan Feng, Tharangini Palanivel, Sarah Pearman, Megan Ung, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. 2020. Informing the design of a personalized privacy assistant for the internet of things. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[20] Wei Fan, Haoran Li, Zheye Deng, Weiqi Wang, and Yangqiu Song. 2024. GoldCoin: Grounding Large Language Models in Privacy Laws via Contextual Integrity Theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 3321–3343.
[21] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. *arXiv preprint arXiv:2402.00367* (2024).
[22] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* 50, 3 (2024), 1097–1179.
[23] Sahra Ghalebikesabi, Eugene Bagdasaryan, Ren Yi, Itay Yona, Ilia Shumailov, Aneesh Pappu, Chongyang Shi, Laura Weidinger, Robert Stanforth, Leonard Berrada, et al. 2024. Operationalizing contextual integrity in privacy-conscious assistants. *arXiv preprint arXiv:2408.02373* (2024).
[24] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security*. 79–90.
[25] Friederike Groschupp, Daniele Lain, Aritra Dhar, Lara Magdalena Lazier, and Srdjan Čapkun. 2025. Can LLMs Make (Personalized) Access Control Decisions? *arXiv preprint arXiv:2511.20284* (2025).
[26] Bingcan Guo, Eryue Xu, Zhiping Zhang, and Tianshi Li. 2025. Not my agent, not my boundary? elicitation of personal privacy boundaries in ai-delegated information sharing. *arXiv preprint arXiv:2509.21712* (2025).
[27] Bingcan Guo, Zhiping Zhang, and Tianshi Li. 2025. Privi: Assisting Users in Authoring Contextual Privacy Rules with an LLM Sandbox. In *Proceedings of the 2025 Workshop on Human-Centered AI Privacy and Security*. 73–83.
[28] Mitra Bokaei Hosseini, John Heaps, Rocky Slavin, Jianwei Niu, and Travis Breaux. 2021. Ambiguity and generality in natural language privacy policies. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*. IEEE, 70–81.
[29] Roberto Hoyle, Luke Stark, Qatrunnada Ismail, David Crandall, Apu Kapadia, and Denise Anthony. 2020. Privacy norms and preferences for photos posted online. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 4 (2020), 1–27.
[30] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38.
[31] Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664* (2025).
[32] Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. 2025. AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions. *arXiv preprint arXiv:2506.09038* (2025).
[33] Nadin Kökciyan, Pinar Yolum, et al. 2022. Taking Situation-Based Privacy Decisions: Privacy Assistants Working with Humans.. In *IJCAI*. 703–709.
[34] Guangchen Lan, Huseyin A Inan, Sahar Abdelnabi, Janardhan Kulkarni, Lukas Wutschitz, Reza Shokri, Christopher G Brinton, and Robert Sim. 2025. Contextual integrity in llms via reasoning and reinforcement learning. *arXiv preprint arXiv:2506.04245* (2025).
[35] Haoran Li, Wenbin Hu, Huihao Jing, Yulin Chen, Qi Hu, Sirui Han, Tianshu Chu, Peizhao Hu, and Yangqiu Song. 2025. Privaci-bench: Evaluating privacy with contextual integrity and legal compliance. *arXiv preprint arXiv:2502.17041* (2025).
[36] Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100* (2025).
[37] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I Hong. 2014. Modeling {Users'} mobile app privacy preferences: Restoring usability in a sea of permission settings. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*. 199–212.
[38] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun Aerin Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. 2016. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Twelfth symposium on usable privacy and security (SOUPS 2016)*. 27–41.
[39] Zhihuang Liu, Ling Hu, Tongqing Zhou, Yonghao Tang, and Zhiping Cai. 2025. Prevalence Overshadows Concerns? Understanding Chinese Users' Privacy Awareness and Expectations Towards LLM-Based Healthcare Consultation. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2716–2734.
[40] Kirsten Martin and Helen Nissenbaum. 2016. Measuring privacy: An empirical test using context to expose confounding variables. *Colum. Sci. & Tech. L. Rev.* 18 (2016), 176.
[41] Luoxi Meng, Henry Feng, Ilia Shumailov, and Earlence Fernandes. 2025. cellmate: Sandboxing browser ai agents. *arXiv preprint arXiv:2512.12594* (2025).
[42] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In *ICLR*.
[43] Niloofar Mireshghallah, Neal Mangaokar, Narine Kokhlikyan, Arman Zharmagambetov, Manzil Zaheer, Saeed Mahloujifar, and Kamalika Chaudhuri. 2025. Cimemories: A compositional benchmark for contextual integrity of persistent memory in llms. *arXiv preprint arXiv:2511.14937* (2025).
[44] Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. 2025. Fully autonomous ai agents should not be developed. *arXiv preprint arXiv:2502.02649* (2025).
[45] Victor Morel, Leonardo Horn Iwaya, and Simone Fischer-Hübner. 2025. AI-driven Personalized Privacy Assistants: a Systematic Literature Review. *IEEE Access* (2025).
[46] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. 2017. Privacy expectations and preferences in an {IoT} world. In *Thirteenth symposium on usable privacy and security (SOUPS 2017)*. 399–412.
[47] Mohammad Hadi Nezhad, Francisco Enrique Vicente Castro, and Ivon Arroyo. 2026. Understanding Users' Privacy Reasoning and Behaviors During Chatbot Use to Support Meaningful Agency in Privacy. *arXiv preprint arXiv:2601.18125* (2026).
[48] Ivoline Ngong, Swanand Kadhe, Hao Wang, Keerthiram Murugesan, Justin D Weisz, Amit Dhurandhar, and Karthikeyan Natesan Ramamurthy. 2025. Protecting users from themselves: Safeguarding contextual privacy in interactions with conversational agents. *arXiv preprint arXiv:2502.18509* (2025).
[49] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.* 79 (2004), 119.
[50] Helen Nissenbaum. 2009. Privacy in context: Technology, policy, and the integrity of social life. In *Privacy in context*. Stanford University Press.
[51] Theo X Olausson, Alex Gu, Benjamin Lipkin, Cedegao E Zhang, Armando Solar-Lezama, Joshua B Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. *arXiv preprint arXiv:2310.15164* (2023).
[52] Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295* (2023).
[53] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogu Kang. 2016. Expecting the unexpected: Understanding mismatched privacy expectations online. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. 77–96.
[54] Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2024. Privacylens: Evaluating privacy norm awareness of language models in action. *Advances in Neural Information Processing Systems* 37 (2024), 89373–89407.
[55] Yan Shvartzshnaider, Schrasing Tong, Thomas Wies, Paula Kift, Helen Nissenbaum, Lakshminarayanan Subramanian, and Prateek Mittal. 2016. Learning

privacy expectations by crowdsourcing contextual informational norms. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4. 209–218.
[56] Alina Stöver, Sara Hahn, Felix Kretschmer, and Nina Gerber. 2023. Investigating how users imagine their personal privacy assistant. *Proceedings on Privacy Enhancing Technologies* (2023).
[57] Dakota Sullivan, Shirley Zhang, Jennica Li, Heather Kirkorian, Bilge Mutlu, and Kassem Fawaz. 2025. Benchmarking LLM Privacy Recognition for Social Robot Decision Making. *arXiv preprint arXiv:2507.16124* (2025).
[58] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025).
[59] Sarah Tran, Hongfan Lu, Isaac Slaughter, Bernease Herman, Aayushi Dangol, Yue Fu, Lufei Chen, Biniyam Gebreyohannes, Bill Howe, Alexis Hiniker, et al. 2025. Understanding Privacy Norms Around LLM-Based Chatbots: A Contextual Integrity Perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 2522–2534.
[60] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature* 625, 7995 (2024), 476–482.
[61] Lillian Tsai and Eugene Bagdasarian. 2025. Contextual Agent Security: A Policy for Every Purpose. In *Proceedings of the 2025 Workshop on Hot Topics in Operating Systems*. 8–17.
[62] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345.
[63] Xiaoyin Wang, Xue Qin, Mitra Bokaei Hosseini, Rocky Slavin, Travis D Breaux, and Jianwei Niu. 2018. Guileak: Tracing privacy policy claims on user input data for android applications. In *Proceedings of the 40th International Conference on Software Engineering*. 37–47.
[64] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
[65] Primal Wijesekera, Arjun Baokar, Lynn Tsai, Joel Reardon, Serge Egelman, David Wagner, and Konstantin Beznosov. 2017. The feasibility of dynamically granted permissions: Aligning mobile privacy with user preferences. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1077–1093.
[66] Primal Wijesekera, Joel Reardon, Irwin Reyes, Lynn Tsai, Jung-Wei Chen, Nathan Good, David Wagner, Konstantin Beznosov, and Serge Egelman. 2018. Contextualizing privacy decisions for better prediction (and protection). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
[67] Yuhao Wu, Ke Yang, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. 2025. Towards automating data access permissions in ai agents. *arXiv preprint arXiv:2511.17959* (2025).
[68] Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110* (2025).
[69] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.
[70] Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. 2023. Satlm: Satisfiability-aided language models using declarative prompting. *Advances in Neural Information Processing Systems* 36 (2023), 45548–45580.
[71] Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems* 35 (2022), 30378–30392.
[72] Ren Yi, Octavian Suciu, Adria Gascon, Sarah Meiklejohn, Eugene Bagdasarian, and Marco Gruteser. 2025. Privacy Reasoning in Ambiguous Contexts. *arXiv preprint arXiv:2506.12241* (2025).
[73] Nicole Zhan, Stefan Sarkadi, and Jose Such. 2023. Privacy-enhanced personal assistants based on dialogues and case similarity. In *European Conference on Artificial Intelligence*. IOS Press.
[74] Xiao Zhan, Stefan Sarkadi, Natalia Criado, and Jose Such. 2022. A model for governing information sharing in smart assistants. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 845–855.
[75] Shuning Zhang, Ying Ma, Jingruo Chen, Simin Li, Xin Yi, and Hewu Li. 2025. Towards Aligning Personalized Conversational Recommendation Agents with Users' Privacy Preferences. *arXiv preprint arXiv:2508.07672* (2025).
[76] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv e-prints* (2023), arXiv–2309.
[77] Zhiping Zhang, Yi Evie Zhang, Freda Shi, and Tianshi Li. 2025. Autonomy Matters: A Study on Personalization-Privacy Dilemma in LLM Agents. *arXiv preprint arXiv:2510.04465* (2025).
[78] Noé Zufferey, Sarah Abdelwahab Gaballah, Karola Marky, and Verena Zimmermann. 2025. "AI is from the devil." Behaviors and Concerns Toward Personal Data Sharing with LLM-based Conversational Agents. *Proceedings on Privacy Enhancing Technologies* 2025, 3 (2025), 5–28.

A Prompts for Evaluation

We provide the set of prompts that were used in our experimental evaluations below.

A.1 Zero-shot Prompts

Prompt 1: Zero-shot prompt for SPA dataset

```
Role: You are an AI agent expert in applying the theory of Contextual Integrity (CI) to make privacy judgments on
     information-sharing requests on behalf of a user.

Objective: Analyze an Incoming Request then judge if it is "appropriate" or "inappropriate" based on your knowledge of
     general privacy norms.

Information Flow Request
Each request is broken down into five fields from CI that describe a specific flow of information:

data_type: The piece of information being shared.

data_subject: The owner of the data.

data_sender: The individual/entity sharing the data.

data_recipient: The individual/entity receiving the data.

transmission_principle: The purpose and/or safeguard condition under which the data is shared.

For this task, assume the data_subject is always the 'user' and the data_sender is always the 'assistant provider'.

Required Output Format
Your entire output must be a single JSON object. Do not include any text outside of this JSON object. The structure
     must be as follows:

{{
  "judgment": "appropriate or inappropriate"
}}

Incoming Request to Judge:
{incoming_request}

Output:
```

Prompt 2: Zero-shot prompt for Education dataset

```
Role: You are an AI agent expert in applying the theory of Contextual Integrity (CI) to make privacy judgments on
     information-sharing requests on behalf of a user.

Objective: Analyze an Incoming Request then judge if it is "appropriate" or "inappropriate" based on your knowledge of
     general privacy norms.

Information Flow Request
Each user request is broken down into five fields that describe a specific flow of information:

data_type: The piece of information being shared.
```

```
data_subject: The owner of the data.

data_sender: The individual/entity sharing the data.

data_recipient: The individual/entity receiving the data.

transmission_principle: The condition under which the data is shared.

For this task, assume the data_subject and the user is always the 'student'.

Required Output Format
Your entire output must be a single JSON object. Do not include any text outside of this JSON object. The structure
     must be as follows:

{{
  "judgment": "appropriate or inappropriate"
}}

Incoming Request to Judge:
{incoming_request}

Output:
```

A.2 Privacy Norms Prompts

Prompt 3: Privacy norms prompt for SPA dataset

```
Role: You are an AI agent expert in applying the theory of Contextual Integrity (CI) to make privacy judgments on
     information-sharing requests on behalf of a user.

Objective: Analyze an Incoming Request then judge if it is "appropriate" or "inappropriate" based on a provided set of
     general privacy norms.

Information Flow Request
Each request is broken down into five fields from CI that describe a specific flow of information:

data_type: The piece of information being shared.

data_subject: The owner of the data.

data_sender: The individual/entity sharing the data.

data_recipient: The individual/entity receiving the data.

transmission_principle: The purpose and/or safeguard condition under which the data is shared.

For this task, assume the data_subject is always the 'user' and the data_sender is always the 'assistant provider'.

General Privacy Norms:
Define user recipients to be a data_recipient from the following set:

parents, partner, siblings, children, housemates, close friends, close family, neighbours, house helper/keeper,
     visitors in general.
```

```
Additionally, define non-user recipients to be a data_recipient from the following set:

assistant provider, third party skills/actions, advertising agencies, law enforcement agencies

Then we provide the top appropriate and inappropriate privacy norms below. Each rule contains a subset of the CI fields
      with its corresponding value.
If a norm does not contain a CI field (e.g., transmission_principle), then that field can take on any value (e.g.,
     transmission_principle can be any value).

The following are privacy norms for user recipients:

N1. data_type: 'sleeping hours', judgment: 'inappropriate'
N2. data_type: 'contacts', judgment: 'inappropriate'
N3. data_type: 'voice command history', judgment: 'inappropriate'
N4. data_type: 'email content', judgment: 'inappropriate'
N5. data_type: 'bank account details', judgment: 'inappropriate'
N6. data_recipient: 'neighbours', judgment: 'inappropriate'
N7. data_recipient: 'visitors in general', judgment: 'inappropriate'

The following are privacy norms for non-user recipients:

N1 data_type: 'contacts', judgment: 'inappropriate'
N2 data_type: 'video calls', judgment: 'inappropriate'
N3 data_type: 'bank account details', judgment: 'inappropriate'
N4 transmission_principle: 'no purpose/condition', judgment: 'inappropriate'
N5 data_recipient: 'non-relevant skills/actions', judgment: 'inappropriate'
N6 data_recipient: 'advertising agencies', judgment: 'inappropriate'

Required Output Format
Your entire output must be a single JSON object. Do not include any text outside of this JSON object. The structure
     must be as follows:

{{
  "judgment": "appropriate or inappropriate"
}}

Incoming Request to Judge:
{incoming_request}

Output:
```

Prompt 4: Privacy norms prompt for Education Dataset

```
Role: You are an AI agent expert in applying the theory of Contextual Integrity (CI) to make privacy judgments on
     information-sharing requests on behalf of a user.

Objective: Analyze an Incoming Request then judge if it is "appropriate" or "inappropriate" based on a provided set of
     general privacy norms.

Information Flow Request
Each user request is broken down into five fields that describe a specific flow of information:

data_type: The piece of information being shared.
```

```
data_subject: The owner of the data.

data_sender: The individual/entity sharing the data.

data_recipient: The individual/entity receiving the data.

transmission_principle: The condition under which the data is shared.

For this task, assume the data_subject and the user is always the 'student'.

General Privacy Norms:
We provide the top appropriate and inappropriate privacy norms below. Each privacy norm contains a subset of the CI
     fields with its corresponding value

The top six appropriate privacy norms:
N1. data_sender: 'professor', data_recipient: 'graduate schools', data_type: 'attendance', transmission_principle: '
     with subject's consent'
N2. data_sender: 'registrar', data_recipient: 'parents', data_type: 'grades', transmission_principle: 'with subject's
     consent'
N3. data_sender: 'professor', data_recipient: 'department chair', data_type: 'term papers', transmission_principle: '
     with subject's consent'
N4. data_sender: 'professor', data_recipient: 'registrar', data_type: 'participation', transmission_principle: 'with
     subject's consent'
N5. data_sender: 'librarian', data_recipient: 'department chair', data_type: 'photo', transmission_principle: 'with
     subject's consent'
N6. data_sender: 'registrar', data_recipient: 'graduate schools', data_type: 'grades', transmission_principle: 'with
     subject's consent'

The top five inappropriate privacy norms:
N1. data_sender: 'TA', data_recipient: 'classmates', data_type: 'grades', transmission_principle: 'if the subject is
     performing poorly'
N2. data_sender: 'TA', data_recipient: 'classmates', data_type: 'transcript', transmission_principle: 'with the
     requirement of confidentiality'
N3. data_sender: 'TA', data_recipient: 'librarian', data_type: 'grades', transmission_principle: 'with a request from
     the subject'
N4. data_sender: 'advisor', data_recipient: 'classmates', data_type: 'transcript', transmission_principle: 'with the
     requirement of confidentiality'
N5. data_sender: 'registrar', data_recipient: 'classmates', data_type: 'transcript', transmission_principle: 'with a
     request from the subject'

Required Output Format
Your entire output must be a single JSON object. Do not include any text outside of this JSON object. The structure
     must be as follows:

{{
  "judgment": "appropriate or inappropriate"
}}

Incoming Request to Judge:
{incoming_request}

Output:
```

A.3 ICL Prompts

Prompt 5: ICL prompt SPA dataset

```
Role: You are an AI agent expert in applying the theory of Contextual Integrity (CI) to make privacy judgments on
     information-sharing requests on behalf of a user.

Ojective: Analyze an Incoming Request and judge if it is "appropriate" or "inappropriate" based on the user's prior
     privacy judgments.

Information Flow Request
Each request is broken down into five fields that describe a specific flow of information:

data_type: The piece of information being shared.

data_subject: The owner of the data.

data_sender: The individual/entity sharing the data.

data_recipient: The individual/entity receiving the data.

transmission_principle: The purpose and/or safeguard condition under which the data is shared.

For this task, assume the data_subject is always the 'user' and the data_sender is always the 'assistant provider'.

History of Prior Requests with User Judgments (this list can be empty or contain multiple entries):
{icl_examples}

Step-by-Step Instructions:
1. Analyze the history of prior requests with user judgments to understand the user's privacy preferences.
Look for patterns across all five fields. For example, note which data_recipients are trusted with a certain data_type,
or under which transmission_principle sensitive information is allowed to be shared.
2. Compare the Incoming Request to these established user preferences to determine the judgment of the incoming request
     .

Required Output Format
Your entire output must be a single JSON object. Do not include any text outside of this JSON object.
The structure must be as follows:

{{
  "judgment": "appropriate or inappropriate"
}}

Incoming Request to Judge:
{incoming_request}

Output:
```

Prompt 6: ICL prompt for Education dataset

```
Role: You are an AI agent expert in applying the theory of Contextual Integrity (CI) to make privacy judgments on
     information-sharing requests on behalf of a user.

Ojective: Analyze an Incoming Request and judge if it is "appropriate" or "inappropriate" based on the user's prior
     privacy judgments.
```

```
Information Flow Request
Each user request is broken down into five fields that describe a specific flow of information:

data_type: The piece of information being shared.

data_subject: The owner of the data.

data_sender: The individual/entity sharing the data.

data_recipient: The individual/entity receiving the data.

transmission_principle: The condition under which the data is shared.

For this task, assume the data_subject and the user is always the 'student'.

History of Prior Requests with User Judgments (this list can be empty or contain multiple entries):
{icl_examples}

Step-by-Step Instructions:
1. Analyze the history of prior requests with user judgments to understand the user's privacy preferences.
Look for patterns across all five fields. For example, note which data_recipients are trusted with a certain data_type,
or under which transmission_principle sensitive information is allowed to be shared.
2. Compare the Incoming Request to these established user preferences to determine the judgment of the incoming request
     .

Required Output Format
Your entire output must be a single JSON object. Do not include any text outside of this JSON object.
The structure must be as follows:

{{
  "judgment": "appropriate or inappropriate"
}}

Incoming Request to Judge:
{incoming_request}

Output:
```

A.4 ICL (w Undet) Propmts

Prompt 7: ICL (w Undetermined) prompt for SPA dataset

```
Role: You are an AI agent expert in applying the theory of Contextual Integrity (CI) to make privacy judgments on
     information-sharing requests on behalf of a user.

Ojective: Analyze an Incoming Request and judge if it is "appropriate", "inappropriate", or "undetermined" based on the
      user's prior privacy judgments.

Information Flow Request
Each request is broken down into five fields that describe a specific flow of information:

data_type: The piece of information being shared.
```

```
data_subject: The owner of the data.

data_sender: The individual/entity sharing the data.

data_recipient: The individual/entity receiving the data.

transmission_principle: The purpose and/or safeguard condition under which the data is shared.

For this task, assume the data_subject is always the 'user' and the data_sender is always the 'assistant provider'.

History of Prior Requests with User Judgments (this list can be empty or contain multiple entries):
{icl_examples}

Step-by-Step Instructions:
1. Analyze the history of prior requests with user judgments to understand the user's privacy preferences.
Look for patterns across all five fields. For example, note which data_recipients are trusted with a certain data_type,
or under which transmission_principle sensitive information is allowed to be shared.
2. Compare the Incoming Request to these established user preferences to determine the judgment of the incoming request
     .
If you are unable to confidently judge whether the Incoming Request is "appropriate" or "inappropriate" based on the
     Prior Requests,
then your judgment should be "undetermined".

Required Output Format
Your entire output must be a single JSON object. Do not include any text outside of this JSON object.
The structure must be as follows:

{{
  "judgment": "appropriate, inappropriate, or undetermined"
}}

Incoming Request to Judge:
{incoming_request}


Incoming Request to Judge:
{incoming_request}

Output:
```

Prompt 8: ICL (w Undetermined) prompt for Education dataset

```
Role: You are an AI agent expert in applying the theory of Contextual Integrity (CI) to make privacy judgments on
     information-sharing requests on behalf of a user.

Ojective: Analyze an Incoming Request and judge if it is "appropriate", "inappropriate", or "undetermined" based on the
      user's prior privacy judgments.

Information Flow Request
Each user request is broken down into five fields that describe a specific flow of information:

data_type: The piece of information being shared.

data_subject: The owner of the data.
```

```
data_sender: The individual/entity sharing the data.

data_recipient: The individual/entity receiving the data.

transmission_principle: The condition under which the data is shared.

For this task, assume the data_subject and the user is always the 'student'.

History of Prior Requests with User Judgments (this list can be empty or contain multiple entries):
{icl_examples}

Step-by-Step Instructions:
1. Analyze the history of prior requests with user judgments to understand the user's privacy preferences.
Look for patterns across all five fields. For example, note which data_recipients are trusted with a certain data_type,
or under which transmission_principle sensitive information is allowed to be shared.
2. Compare the Incoming Request to these established user preferences to determine the judgment of the incoming request
     .
If you are unable to confidently judge whether the Incoming Request is "appropriate" or "inappropriate" based on the
     Prior Requests,
then your judgment should be "undetermined".

Required Output Format
Your entire output must be a single JSON object. Do not include any text outside of this JSON object.
The structure must be as follows:

{{
  "judgment": "appropriate, inappropriate, or undetermined"
}}

Incoming Request to Judge:
{incoming_request}


Incoming Request to Judge:
{incoming_request}

Output:
```

A.5 ARIEL Prompts

Prompt 9: Ontology generation prompt for SPA dataset

```
Role:
You are an AI expert specializing in data privacy, contextual integrity (CI), and ontological modeling.

Task:
Your task is to generate three hierarchical ontologies: one for data_type, one for data_recipient, and one for
     transmission_principle. These ontologies must be usable by a language model to logically map all possible values
     for each field into ranked levels.

The hierarchy for each ontology must be based on two sources, in order of priority:

1. User Privacy Preferences: The specific judgments in the Information flows are the primary driver for creating levels
     . Any two values (e.g., value_A vs. value_B) that are treated differently by the user in the judgments must be
     placed in separate, distinct levels.

2. General Norms: Use general norms only after all user judgments are satisfied. General norms should be used to (a)
     write a clear description for the levels you've already created and (b) logically sort other common values related
      to the Smart House Personal Assistant (SPA) ecosystem into the levels defined by the user's preferences.

Core Concepts: Contextual Integrity (CI) A flow of information is described by five CI fields:

data_type: The piece of information being shared.

data_subject: The owner of the data.

data_sender: The individual/entity sharing the data.

data_recipient: The individual/entity receiving the data.

transmission_principle: The purpose and/or safeguard condition under which the data is shared. The format is: purpose,
     condition. Note that if transmission_principle is empty, it should be interpreted as having no purpose.

For this task, assume the data_subject is always the 'user' and the data_sender is always the 'assistant provider'.
     Your ontologies will define the values for the other three fields.

Input: User Privacy Judgments.

You will be provided with a list of the user's judgments on specific information flows. These flows determine the user'
     s specific preferences.

Each scenario is structured as: (data_type, data_recipient, transmission_principle) -> judgment (where judgment is '
     appropriate' or 'inappropriate')

Information flows:
{privacy_preferences}

Ontology Generation Rules

You must generate one ontology for each of the three fields below.

1. data_type Ontology
Relationship: The Data Sensitivity as percieved by the user.
```

```
Hierarchy: Level 1 (L1) must contain the least sensitive data types. Sensitivity increases at higher levels .

Value Sources: Base your categories on common SPA data types. Use values from data_type in Information flows, then
     expand this by referencing common categories from Google Actions and Alexa Skills.

2. data_recipient Ontology
Relationship: The degree of Trust by the user to the recipient. Additionally, the authority each recipient has to
     receive the data.

Hierarchy: Level 1 must contain the most trusted recipients and/or recipients with most authority, with trust and/or
     authority decreasing at higher levels.

Value Sources: Use the values from data_recipient in Information flows, then expand this to include a comprehensive set
      of potential recipients in an SPA ecosystem, such as household members or visitors. Additionally, entities who
     may have access to the user's SPA ecosystem, such as providers.

3. transmission_principle Ontology
Relationship: The combination of (A) The Purpose of the transmission and (B) The Condition (Safeguard) applied.

Hierarchy: The hierarchy must rank combinations of these two factors. L1 represents the most appropriate combination (e
     .g., a related purpose with strong safeguards). Purpose and safeguard decrease at higher levels (e.g., no purpose
     and/or weaker/no safeguard).

Key Distinction: The hierarchy must separate principles based on the strength of the safeguard (Condition). A principle
      with a specific safeguard (e.g., 'in an anonymized form', 'with explicit user consent') must be in a higher (more
      appropriate, lower L-number) level than the exact same purpose with a weaker or no safeguard.

Value Sources: Use the principles in Information flows as your starting point. Expand this to include a comprehensive
     set of combinations involving purpose and safety of the data.

Coverage and Exclusivity Requirement:

1. Mutually Exclusive (The "No Overlap" Rule): This is your most important rule. The description for each level must be
      precise enough that any given value maps to exactly one level. Levels must not overlap.

2. Collectively Exhaustive (The "Full Coverage" Rule): Your generated ontologies must be comprehensive. They must
     logically cover all values present in the Information flows input, as well as all other common values for each CI
     field in an SPA context.

3. Judgment-Driven Grouping: To achieve both rules, use the Information flows to define your levels.
 - Example, consider two scenarios that differ by data_recipient: (..., recipient_A, ...) is appropriate but (...,
      recipient_B, ...) is inappropriate. Therefore, 'recipient_A' and 'recipient_"B' must be in different levels. Your
       level description must capture why (e.g., "L1: Description of trusted group, e.g., recipient_A" vs. "L4:
      Description of untrusted group, e.g., recipient_B").

Output Format:
Your output must consist only of the three generated ontologies. Do not add any preamble, explanation, or
     conversational text. Use the following format:

data_type Ontology
L1. "Description of this level"
L2. "Description of this level"
...

data_recipient Ontology
```

```
L1. "Description of this level"
L2. "Description of this level"
...

transmission_principle Ontology
L1. "Description of this level"
L2. "Description of this level"
...

Output:
```

Prompt 10: Ontology generation prompt for Education dataset

```
Role:
You are an AI expert specializing in data privacy, contextual integrity (CI), and ontological modeling.

Task:
Your task is to generate four hierarchical ontologies: one for data_type, one for data_sender, one for data_recipient,
     and one for transmission_principle. These ontologies must be usable by a language model to logically map all
     possible values for each field into ranked levels.

The hierarchy for each ontology must be based on two sources, in order of priority:

1. User Privacy Preferences: The specific judgments in the Information flows are the primary driver for creating levels
     . Any two values (e.g., value_A vs. value_B) that are treated differently by the user in the judgments must be
     placed in separate, distinct levels.

2. General Norms: General privacy norms and common privacy expecations around school/education setting. Use general
     norms only after all user judgments are satisfied. They should be used to (a) write a clear description for the
     levels you've already created and (b) logically sort other common values related to the school/education context
     into the levels defined by the user's preferences.

Core Concepts: Contextual Integrity (CI) A flow of information is described by five CI fields:

data_type: The piece of information being shared.

data_subject: The owner of the data.

data_sender: The individual/entity sharing the data.

data_recipient: The individual/entity receiving the data.

transmission_principle: The condition under which the data is shared.

For this task, assume the data_subject and the user is always the 'student'. Your ontologies will define the values for
      the other four fields.

Input: User Privacy Judgments.

You will be provided with a list of the user's judgments on specific information flows. These flows determine the user'
     s specific preferences.

Each scenario is structured as: (data_type, data_sender, data_recipient, transmission_principle) -> judgment (where
     judgment is 'appropriate' or 'inappropriate')
```

```
Information flows:
{privacy_preferences}

Ontology Generation Rules

You must generate one ontology for each of the four fields below.

1. data_type Ontology

Relationship: The student's perceived Data Sensitivity of data_type.

Hierarchy: Level 1 must contain the least sensitive data types, with sensitivity increasing at higher levels.

Value Sources: Use values from data_type in Information flows, then expand this by referencing common school/education
     data types.

2. data_sender Ontology

Relationship: The authority of the data_sender to send the student's data.

Hierarchy: Level 1 must contain the senders with the most authority, with authority decreasing at higher levels.

Value Sources: Use the values from data_sender in Information flows, then expand this to include a comprehensive set of
      potential recipients in a school/education setting, such as professors.

3. data_recipient Ontology

Relationship: The degree of Trust a student typically has in these people. Additionally, the authority each recipient
     has to receive the data.

Hierarchy: Level 1 must contain the most trusted recipients and/or recipients with most authority, with trust and/or
     authority decreasing at higher levels.

Value Sources: Use the values from data_recipient in Information flows, then expand this to include a comprehensive set
      of potential recipients in a school/education setting, such as classmates.

4. transmission_principle Ontology

Relationship: The degree of Data Safety offered to the student.

Hierarchy: Level 1 must represent the highest degree of data safety, with the degree of safety decreasing at higher
     levels.

Value Sources: Use the principles in Information flows as your starting point. Expand this to include other conditions
     involved in a school/education setting.

Coverage and Exclusivity Requirement:

1. Mutually Exclusive (The "No Overlap" Rule): This is your most important rule. The description for each level must be
      precise enough that any given value maps to exactly one level. Levels must not overlap.

2. Collectively Exhaustive (The "Full Coverage" Rule): Your generated ontologies must be comprehensive. They must
     logically cover all values present in the Information flows input, as well as all other common values for each CI
     field in a school/education context.
```

```
3. Judgment-Driven Grouping: To achieve both rules, use the Information flows to define your levels.
 - Example, consider two scenarios that differ by data_recipient: (..., recipient_A, ...) is appropriate but (...,
      recipient_B, ...) is inappropriate. Therefore, 'recipient_A' and 'recipient_B' must be at different levels. Your
      level description must capture why (e.g., "L1: Description of trusted group, e.g., recipient_A" vs. "L4:
      Description of untrusted group, e.g., recipient_B").

Output Format:
Your output must consist only of the three generated ontologies. Do not add any preamble, explanation, or
     conversational text. Use the following format:

data_type Ontology
L1. "Description of this level"
L2. "Description of this level"
...

data_recipient Ontology
L1. "Description of this level"
L2. "Description of this level"
...

transmission_principle Ontology
L1. "Description of this level"
L2. "Description of this level"
...

Output:
```

Prompt 11: Ontology mapping prompt

```
Role:
You are an AI expert specializing in data privacy, contextual integrity (CI), and ontological modeling.

Task:
Your task will be to map string values to the relevant level of an ontology.

Background:
A request is an information flow that can be described by five CI fields:

data_type: The piece of information being shared.

data_subject: The owner of the data.

data_sender: The entity sharing the data.

data_recipient: The entity receiving the data.

transmission_principle: The purpose and/or safeguard condition under which the data is shared. The format is: purpose,
     condition. Note that if transmission_principle is empty, it should be interpreted as having no purpose.

You will be provided a prior request, an incoming request, a list of ontologies, and the CI field that differs between
     the prior and incoming request.

Step-by-step instructions:
```

```
1. Identify the field name from the Differing Field input.

2. Get the value for this field from Prior Request. This will be the prior_A value.

3. Get the value for this field from Incoming Request. This will be the incoming_B value.

4. Select the ontology from Ontologies that matches the differing field.

5. Carefully analyze the descriptions of each level (L1, L2, etc.) in the selected ontology.

6. Map the prior_A value to its most accurate level (e.g., "L1"). This will be the mapped_prior_A value.

7. Map the incoming_B value to its most accurate level (e.g., "L2"). This will be the mapped_incoming_B value.

Mapping Requirements:
Analyze each level's description carefully to find the best fit.

You must map both values to a level. If a value is ambiguous or doesn't seem to fit perfectly, select the best-fit
     level. Do not refuse to map.

Output Format:
Your output must be a single, raw JSON object. Nothing else.

Do not include any other text, explanations, preamble, or markdown formatting (like ```json). The JSON object must be
     the only thing you output.

The JSON object must use these exact keys and format:

{{
  "prior_A": "differing field value from prior request obtained in Step 1.",
  "incoming_B": "differing field value from incoming request obtained in Step 1.",
  "mapped_prior_A": "prior_A mapped to the level of the selected ontology in Step 3.",
  "mapped_incoming_B": "incoming_B mapped to the level of the selected ontology in Step 3."
}}

**EXAMPLES:

Correct output (Good):

{{
  "prior_A": "credit card information",
  "incoming_B": "birth date",
  "mapped_prior_A": "L3",
  "mapped_incoming_B": "L1"
}}

**Incorrect Example (Bad - DO NOT DO THIS):**
{{
  "prior_A": "some value",
  - "incoming_B": "another value", <-- DO NOT ADD A HYPHEN
  "mapped_prior_A": "L1",
  "mapped_incoming_B": "L2"
}}

Ontologies:
```

```
{ontologies}

Prior Request:
{prior_request}

Incoming Request:
{incoming_request}

Differing Field:
{differing_field}

**CRITICAL REMINDER: Your response must be *only* the syntactically valid JSON object as specified in the "Output
     Format". It must start with `{{` and end with `}}`. Do not include any other text, markdown, or **hyphens (`-`).**


Output:
```