

Towards Automating Data Access Permissions in AI Agents

Yuhao Wu*, Ke Yang†, Franziska Roesner‡, Tadayoshi Kohno§, Ning Zhang*, Umar Iqbal*

*Washington University in St. Louis †University of California, Irvine ‡University of Washington §Georgetown University

Abstract—As AI agents attempt to autonomously act on users’ behalf, they raise transparency and control issues. We argue that permission-based access control is indispensable in providing meaningful control to the users, but conventional permission models are inadequate for the automated agentic execution paradigm. We therefore propose *automated permission management for AI agents*. Our key idea is to conduct a user study to identify the factors influencing users’ permission decisions and to encode these factors into an ML-based permission management assistant capable of predicting users’ future decisions. We find that participants’ permission decisions are influenced by communication context but importantly individual preferences tend to remain consistent within contexts, and align with those of other participants. Leveraging these insights, we develop a permission prediction model achieving 85.1% accuracy overall and 94.4% for high-confidence predictions. We find that even without using permission history, our model achieves an accuracy of 66.9%, and a slight increase of training samples (i.e., 1–4) can substantially increase the accuracy by 10.8%.

1. Introduction

LLMs have enabled a new computing paradigm, in which the system (referred to as an *AI agent* or an *agentic system*) relies on machine learning models to autonomously resolve user queries expressed in natural language [76]. For example, to resolve a user query to *book a flight*, the AI agent might engage with the necessary tools (e.g., a travel reservation tool), automatically access the required user information (e.g., from the system storage/memory), use the user’s credit card details, and make the booking. While this execution paradigm is tremendously powerful and is enabling exciting use cases [7], [25], [61], there are serious security and privacy risks to consider.

At a high level, a key concern is that the AI agent’s actions may not align with the user’s intentions or expectations, which could be due to untrustworthy system modules influencing the LLM [33], [78], [85], LLM making mistakes on its own [31], [44], [82], or the user simply not agreeing with the agent’s course of action [41], [54], [75]. These concerns apply to a broad range of the agent’s activities, such as autonomous data access and actions in the real-world. In this paper, *we focus on the security and privacy issues pertaining to the access of data*. For example, we attempt to limit the unexpected sharing of a user’s credit card details with the wrong tool.

Fundamentally, the key issues are limited user transparency and control over the data access by AI agents. Prior computing systems, such as mobile platforms, have encountered similar problems and have relied on permission models (in which users are asked to grant applications permission to access sensitive resources) to provide users transparency and control over data usage in the system [21], [22], [26], [67], [68]. Similarly, AI agents will need permission management models to control access to resources.

Permission management, however, needs to be designed anew for AI agents, where new challenges arise that stretch the limits of existing permission models. For example, as AI agents generate new functionalities based on input from various system modules, data needed to resolve user queries may not be known beforehand, thus diminishing the utility of legacy install time permissions. Similarly, as AI agents may need several pieces of user information to resolve a user query, constantly interrupting users for permissions is incompatible with the automation promised by AI agents, thus diminishing the utility of standard runtime permissions.

Considering that *automation* is a key value proposition of AI agents, *a permission management system that can automatically make decisions on users’ behalf is critical for AI agents*. Building an automated permission management requires addressing two important problems. First, understanding the factors that users consider or factors that otherwise influence users’ permission decision-making. Second, making permission decisions that precisely meet user needs and expectations for future data sharing, including previously unseen data types. This paper takes a multifaceted approach, including both (i) conducting a user study to understand user permission preferences and (ii) exploring the design of a permission prediction system along with its implementation and evaluation.

To understand users’ data sharing permission preferences, we conduct a vignette-based user study. We create a bespoke user study setup by developing our own website that attempts to immerse participants in training a futuristic personal assistant, including training it to share data, as per their needs. We capture differing user preferences across a wide range of questions spanning several domains, such as health and fitness, finance, and entertainment. Our findings indicate that fewer participants express permissions to AI agents for automatic enforcement when they make mistakes, and participants often struggle with providing appropriate permissions, i.e., under- and over-permissioning are common issues in agentic systems, similar to prior systems [43], [63], [70]. We also note that participants’ permission deci-

sions are influenced by the context of the communication, choice of data being shared, and their privacy consciousness. Importantly, we find that, at least in the context of our study, participants’ permission preferences remain consistent, are similar across various communication contexts, and are often similar to other participants—presenting an opportunity to predict their future permissions.

To learn and predict users’ data sharing permission preferences, we explore developing a hybrid machine learning framework that learn from individual user preferences and also from preferences of similar users (using data from our user study). We leverage LLM in-context learning [57] to learn individual user preferences for two crucial reasons: (i) we possess limited permission decision history for each user, and LLMs have demonstrated attaining high accuracy with limited training data, i.e., “few shots” [19], [57]; (ii) permission decisions need to be made for unseen data, which LLMs allow to make without retraining [13]. We leverage collaborative filtering [29], [30] because it allows us to learn from the preferences of other similar users. Our final resulting model combines both in-content learning and collaborative filtering, as they complement each other. We achieve an accuracy of 85.1% (with a recall of 85.2% and a precision of 92.8%). We also explore adjusting prediction confidence score thresholds and find that a stricter threshold, we can achieve an accuracy of 94.4%, but compromise on making predictions for 74.1% of the data. We also observe that even a slight increase in training data (i.e., user permission decision history), our classification accuracy substantially increases. For example, accuracy reaches 66.9% without permission history, but incorporating history from just 1–4 queries improves it by 10.8%. To foster future research, we release our user study data and code¹.

Our key contributions are as follows:

- 1) We propose automating data access permissions in AI agents, such that a *permission assistant* can observe a user’s permission decision history and can make automatic decisions on the user’s behalf in the future.
- 2) To realize our goal of automating permission decisions, we develop a bespoke vignette-based user study to understand various factors that may influence users’ data-sharing permission decisions in AI agents. We then conduct the study with 205 participants on Prolific.
- 3) We translate the insights from our user study into a permission inference framework capable of predicting users’ permission preferences, achieving 85.1% accuracy overall and 94.4% for high-confidence predictions.

2. Background and Motivation

2.1. AI Agents

While there is no standardized AI agent architecture, at their core, AI agents (often also referred to as *agentic systems*) consist of an LLM (typically accessed via an API), a system prompt (that defines the agent’s functionality),

memory (to keep a record of user interactions and data), and a set of tools (to take action in the real world) [76]. To resolve a user query, AI agents identify the relevant tools and data, formulate an execution plan (i.e., a set of instructions) for an LLM, and autonomously act on the formulated plan.

For example, for a user query to “book a flight”, the agent will first determine that it needs to use a travel reservation tool and requires the user’s information (such as the user’s name and date of birth). The execution plan may include instructions to search for flights, provide payment details, and make the booking. Acting on the plan will include the agent calling the appropriate travel reservation tool APIs with the relevant data, listed in the execution plan.

AI agents exhibit varying degrees of autonomy; in some cases, they can complete tasks entirely of their own, while in other cases they require some level of human supervision during execution. For example, for the flight booking request, agents may retrieve trip dates and locations from the user query, the user’s name and date of birth from the memory storage, and may only request the user to provide credit card information.

2.2. Security and Privacy Risks

As the automated execution paradigm of AI agents makes user interactions seamless, there are obvious benefits to it; however, it also presents serious security, privacy, and safety issues to the users. At a high level, a key concern is that the AI agent’s actions may not align with the user’s intentions or expectations. The misalignment could be due to a variety of reasons, such as untrustworthy system modules influencing the LLM, LLM making mistakes of its own, or the user simply not agreeing with the agent’s course of action. These concerns apply to a broad range of agents’ activities, such as agents using user data of their own or taking actions on users’ behalf in the real world, which are distinct and require tailored treatment. For the scope of this paper, we only focus on limiting the security and privacy issues that pertain to the usage of data.

Next, we describe some of the fundamental issues that could lead to a misalignment between users’ expectations and AI agents’ data usage practices.

Inherent Limitations in Agentic Execution Paradigm.

AI agents enhance their capabilities through exposure to system resources, such as data and tools. It means that only when an agent is aware of the system capabilities (i.e., a particular tool or a piece of data) can it use those. Thus, agents are often designed to get unrestrained access to system resources, including the ones that may not be needed to address the user query, which violates the principle of least privilege. As LLMs are susceptible to prompt injection, malicious resources (such as malicious third-party tools or files) can exploit the LLM’s unrestrained system access to read sensitive user data or influence the LLM’s execution.

Another key issue is that the LLM’s interfacing is based on natural language, which can be imprecise and ambiguous [34], [45]. It means that even in non-adversarial

1. <https://github.com/llm-platform-security/ai-agent-permissions>

scenarios, the underlying LLM in an AI agent might collect incorrect or unnecessary user data. For example, if the travel reservation tool imprecisely specifies that it needs *relevant user data* to make a travel reservation, the LLM’s interpretation of this data may be different than that of the tool and result in inadvertent collection and sharing of unnecessary user data. Hallucination issues in LLMs may further exacerbate these problems [34].

Untrustworthy Third-party Tools. AI agents rely on input from several system modules to determine and steer their execution. Some of these modules, chiefly tools, are developed by third-party developers and load unvetted content from the internet, which makes them an unreliable source to determine the system’s execution. For example, a malicious, compromised, or buggy travel reservation tool may direct the agent to include users’ passport numbers for booking flights even when it is not necessary (such as for domestic flights). Prior research has already shown that third-party tools often collect more data than is needed, including sensitive data prohibited by the platforms [78]. While the problem of excessive data collection by third-party services has existed in prior computing platforms [20], [32], [65], [74], it presents elevated risks in the case of AI agents, because of their pivotal reliance on third-party tools to determine their execution.

Users’ Privacy Consciousness. Users’ expectations, desires, and privacy needs may also simply not align with an AI agent’s action. For example, prior research has shown that users often feel uneasy about sharing intimate data (e.g., health, finances) with AI assistants, fearing misuse or eavesdropping [40], [49]. Prior research also shows that users differ significantly in their willingness to share data, which is often influenced by context trust, tech literacy, and prior experience [2], [8], [49]. In some cases, users may refuse to avail some services or compromise on the user experience. For example, in the context of the flight booking example, users may prefer providing the origin flight city manually, instead of having an agent infer it from the device’s GPS sensor.

2.3. Permission-Based Data Access

Fundamentally, the key issues are limited transparency and control over the data access by AI agents. Prior computing systems and platforms, such as mobile platforms, have encountered similar problems and have relied on *permission models* to provide user transparency and control over the data usage in the system [21], [22], [26], [67], [68]. Likewise, agents can also benefit from a permission management model to control access to resources.

Deployed AI agents, such as ChatGPT, currently adapt permission models from conventional systems [58]; however, as we describe next, they fall short in supporting and instead hinder the automated agentic execution paradigm.

Insufficiency of Existing Permission Models. Permission management systems for AI agents need to be developed anew as they significantly differ from conventional comput-

ing systems. Specifically, the install-time [5] and runtime [6] permissions from conventional systems are insufficient, especially considering the automated execution paradigm of AI agents. For example, install-time permissions assume that the resources needed by applications are known beforehand, and thus users can make permission decisions at the time of installation. In contrast, in most cases of agent execution, behavior is determined at runtime based on input from system modules, which can lead to the emergence of new behaviors. Thus, the resources (which can be very broad) needed to solve a query may not be known beforehand [76], [81]. Similarly, runtime permissions (upon which smartphone platforms have largely converged) allow users to manually make decisions on access of individual resources, which suits conventional systems as only a handful of resources are accessed at runtime (and where user interactions already follow a typical UI flow), such as granting location access permission in mobile platforms. Whereas agents often require accessing several pieces of user data at runtime to solve a query, thus merely applying legacy runtime permission models can degrade the user experience and contribute to permission fatigue.

While conventional permission models are mostly limited in supporting the agentic execution paradigm, they could still be suitable for some use cases. For example, AI agents could leverage install-time permissions to manage OAuth-based authentication [1] in AI agents.

3. Towards Automated Permissions Management in AI Agents

Considering that a significant number of data resources are accessed at runtime to facilitate execution of queries—more than users can reasonably be asked to constantly evaluate and decide upon—we argue that a permission management system that can automatically make decisions on users’ behalf is a necessity for agents. Our observation is consistent with previous work on voice-based personal assistants, which found that while users want control, they dislike excessive prompts and prefer automated permission management with minimal interruptions [49], and permission prompt fatigue has been a longstanding known issue in other contexts as well [11], [22], [53], [72].

3.1. Research Goals

Developing automated permission management requires addressing three key challenges: (i) understanding diverse user data sharing preferences, (ii) accurately learning and predicting user preferences, and (iii) reliably enforcing predicted preferences. In this paper, we focus on (i) and (ii); prior complementary work [8], [52], [79] can be used to address (iii). We describe our approaches to achieve these goals below.

3.1.1. Goal 1: Understanding diverse user preferences. A prerequisite to making permission decisions on users’

behalf is to understand the preferences and expectations of a variety of users. To understand user preferences, we conduct a vignette-based user study, which presents several scenarios to participants to capture their preferences. Building on the prior work, we identify several factors, such as the context of the communication, privacy consciousness, and users’ prior experiences, that may influence user preferences [2], [8], [40], [49]. Our user-study presents several scenarios to users to capture their preferences across a wide range of factors that may influence their preferences.

While prior work exists on understanding user preferences and expectations for personal assistants, it mostly focused on older non-LLM technologies [2], [40], [47], [48], [49]. Newer LLM-based personal assistants/agents are fundamentally different as they rely on a new natural language-based automated execution paradigm and offer far more advanced capabilities. For example, non-LLM agents mostly supported single-dialog user interactions and carried out limited tasks automatically; consequently, prior studies only analyzed a handful of scenarios while examining user permission preferences [49]. Whereas LLM-agents support multi-dialog user interactions and carry out many tasks automatically [50], [81]. Users are likely to form different mental models for modern LLM-based AI agents than for traditional, non-LLM agents, so tailored user studies are essential to capture these distinct experiences. To the best of our knowledge, we are the first to study user preferences in automating data sharing permissions in the context of LLM-based AI agents. Our goal is also not just to simply understand user preferences, but to translate them to a system design and explore their utility in automatically predicting permission decisions. Prior work has also not explored understanding user permission preferences in AI-based systems for the purposes of training an automatic permission prediction assistant.

3.1.2. Goal 2: Accurately learning user preferences. To enforce user preferences across a wide spectrum of conversational contexts and a range of data types, it is crucial to develop a permission preference prediction system that accurately learn users’ preferences from a handful of contexts and applies to other unseen contexts. More specifically, in a real world setting, users may only provide preferences for a select few scenarios, and the system may encounter scenarios that it has not seen before. To tackle these challenges, we rely on a combination of collaborative filtering [29] and LLM-based in-context learning [57] to develop our permission inference system. We choose collaborative filtering because it enables learning a user’s preferences by analyzing the preferences of similar users [29], [30], thus avoiding the data sparsity issues. We choose LLM in-context learning, as it can allow the systems to continuously refine permission inference and predict new previously unseen data types, without requiring re-training, as new contexts and data types are seen [19], [57].

Prior work has proposed predicting user permissions in the context of mobile [12], [23], [46], [51], [80], IoT [9], [15], [71], and non-LLM personal assistants [4], [83], [84].

However, prior work can only predict a handful of *known* system resources and data types, which suited the needs of older systems. For instance, notable studies on non-LLM agents/assistants manage data access permissions across only 15 data types at a coarse granularity [4], [83], [84]. In contrast, LLM-based AI agents routinely encounter new previously unseen data types as users explore new use cases [78], and thus prior approaches simply cannot scale to AI agents. For example, a classic ML model trained on a set of data types used by a tool/app would require retraining to make predictions for new unseen data types used by the tool/app. Our proposed approach tackles this challenge by relying on LLM in-context learning, which does not require retraining to make permission decisions on unseen data.

3.2. Threat Model

System Model. We assume AI agents that maintain user collected data in dedicated *memory modules* and attempt to automatically use that data to provide personalized services to the users. These AI agents include well-known personal assistants, such as OpenAI’s ChatGPT [60], as well as AI agents developed through agent development toolkits, such as LangChain [28] and LlamaIndex [35]. These AI agents also support third-party tools, and both the AI agent and third-party tools can collect and use user data. AI agents also include system scaffolding that allows them to control the execution flow, e.g., accessing data from memory, initiating LLMs, and making network requests via tools [76].

Adversaries and Goals. We assume that the third-party tools could be untrustworthy/dishonest, malicious, or compromised (e.g., via a prompt injection). The attacker’s goals are to leverage third-party tools to steal sensitive data that is present in the agent’s memory or exists with other tools. We assume the LLM to be not malicious but error-prone and making mistakes in accessing unnecessary and sensitive data, e.g., due to ambiguity of natural language [34], [45].

Trust Relationships. We assume that the AI agent and its scaffolding to be trustworthy, and do not have any direct intent to harm users (though they are still vulnerable to attacks, e.g., prompt injection). We also assume that the users may not share data with the AI agents and tools when they think it is unnecessary or simply do not feel comfortable sharing that data, despite data being necessary for the AI agents and tools to provide functionality.

Permission Assistant Design Assumptions. To understand user preferences, we communicate to users (in our user study) that the tools may be malicious, compromised, or dishonestly collect more data than they need. We envision the deployment of our permission assistant in the system scaffolding of the AI agent, without the LLM having direct access to the permission assistant. As the permission assistant makes probabilistic decisions, it may unintentionally make incorrect predictions in some cases. However, the permission assistant does not act with malicious intent, and users retain control over whether to accept its decisions.

Based on these considerations, we assume the permission assistant to be trustworthy.

Out of scope. As the implementation and deployment of the permission assistant requires solving unique challenges of its own, we do not consider it in the scope of this paper. For example, a secure personal assistant may require sandboxed or TEE-based deployment, detection of malicious data flows, and control of the generation of LLM instructions. We envision that the prior work on these topics in the context of AI-agents (e.g., [8], [16], [52], [79]) can be extended to implement a secure permission assistant.

4. Understanding User Preferences

To automatically make permission decisions on users' behalf, it is crucial to first understand how users make permission decisions in the context of agentic systems. To this end, we conduct a user study to understand users' permission decision process. Our goal with the user study is to understand factors that influence users' decision making, so that they can help inform the design of automated permission management systems.

4.1. Study Design

4.1.1. Overview. We consider several variables in our user study (such as demographics, privacy consciousness, usage context) that prior research has identified to influence users' permission decisions in other computing platforms [2], [8], [49]. Meanwhile, AI agents introduce unique capabilities that users have not experienced in prior systems. To help users become familiar with these capabilities and their associated risks, and to examine the influence of all factors in a controlled setting, we develop a vignette-based study [24], which has been used in prior work investigating users' privacy preferences [27], [38], [39], [42], [55], [69]. In our study, we present vignettes to users that attempt to immerse the participants in training a futuristic personal assistant to their needs. Specifically, we present scenarios to participants, where they are asked to help the personal assistant: (i) pick the right set of tools to solve a query, (ii) pick the right set of data to solve a query, and (iii) automatically make data accessing and sharing permission decisions on their behalf. We present a total of 5 questions for tool and data selection, and 20 questions for permission decisions.² To investigate how potential adversarial manipulations or mistakes may influence users' decisions, for 25% (5) of permission decision questions, we include incorrect (unnecessary) data types and ask participants to express their data sharing permission preferences. (Our analyses in Section 4.2.1 and 4.2.2 consider both incorrect and correct data, while the remaining sections only consider correct data.) We also include a warning for all permission decision

2. We present 4 options for permission decisions: (1) Yes, always share, (2) Yes, but ask me next time, (3) No, but ask me next time, and (4) No, never share. These options are the same as the permission options used by OpenAI's custom GPTs [59] (except for never share, which is not an option in OpenAI's ecosystem).

questions that prompts the participants to be careful, as the agent or tools may collect incorrect or unnecessary data. Additionally, we develop our own custom website to make the experience more immersive for users³.

4.1.2. Question curation. We develop an LLM-based framework to curate questions for our user study. We begin by selecting a set of 8 domains⁴ and curating 21 tools (spanning all domains) to record participant preferences on a variety of topics. Our tool curation involves listing functionalities offered by the tools and the data needed to provide those functionalities, leveraging prior work on tool curation [17], [78]. We treat the data needed by tools as ground truth.

We iteratively provide domains and their tools to the LLM, and prompt it to generate queries that users might ask an AI agent, and that require using one or more tools. Once queries are generated, three members of our research team review the generated queries to filter out redundant queries and semantically group similar data types (e.g., address and location). In summary, we curate 65 questions⁵ spanning 8 domains, involving 142 unique data types and 75 generic data types across 21 tools.

4.1.3. Priming participants. Before participants pursue the questions in the user study, we prime the participants by asking them questions about their privacy consciousness (e.g., the importance of privacy to the participants, participants' key privacy concerns). Prior research has shown that asking about privacy attitudes at the start of a study can raise participants' privacy awareness and prompt more cautious behavior throughout the task [18], [66], [73]. Our goal with priming is to make participants privacy conscious, so that their choices more accurately reflect their privacy posture, as it may in real life when they are training an assistant to make permission decisions on their behalf.

4.1.4. Participant recruitment. We recruit a total of 205 participants (we remove 2 participants because of invalid responses) from Prolific [64] in the US, with a minimum Prolific prior survey approval rate of 90% and age over 18 years. Our study takes approximately 18 minutes to complete, and we pay \$4.20 to each participant (using \$14 as the hourly wage at the lead institution). Our study was reviewed by our institution's review board (IRB) and deemed exempt.

4.2. Findings

4.2.1. When AI agents make mistakes, fewer participants express data sharing permissions, but more participants express not sharing permissions. As we explore automating users' data sharing permissions decisions, we

3. We present the screenshots of our user study website at <https://github.com/llm-platform-security/ai-agent-permissions/blob/main/website.pdf>.

4. Entertainment, Health & Fitness, Smart Home, Travel, Shopping, Work & Productivity, Social, and Finance.

5. The questions are detailed at <https://github.com/llm-platform-security/ai-agent-permissions/blob/main/queries.json>.

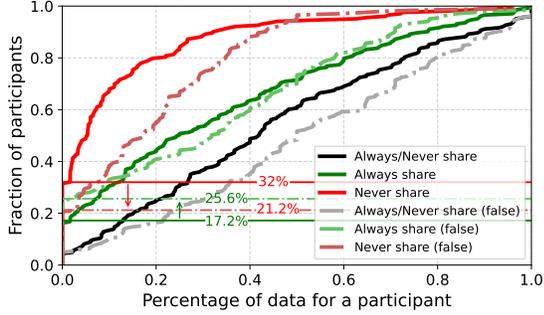


Figure 1: Distribution of participant permission preferences. The *false* label indicates data sharing instances where unnecessary data types were presented to participants (along with the necessary data types).

first investigate permission preferences that participants expressed to the AI agent. We consider *yes*, *always share* and *no*, *never share* permissions, as permissions expressed for the AI agent to learn participant preferences. Figure 1 presents a distribution of participants and their permission preferences. We note that 95.1% of the participants express *always share* permission decisions to AI agents for one or more data sharing decisions. The proportion of permission decisions is higher for sharing data than for not sharing data, which suggests that users may engage in over-permissioning (discussed in Section 4.2.2). Specifically, 82.8% and 68.0% of participants let AI agents to automatically share and not share data at least once, respectively.

Since real-world AI agents can make mistakes and over-collect data, we next explore how user permission preferences change when AI agents make mistakes. Dotted lines in Figure 1 present distributions of participants and their permission preferences when the AI agent makes mistakes. We find that participants tend to express more permission decisions for not sharing data (i.e., *never share*). Specifically, the number of participants who never express *never share* permission decision decreases from 32% (when participants are presented with necessary data) to 21.2% (when participants are presented with unnecessary data, along with necessary data). This suggests that AI agent mistakes prompt users to assess their permission decisions, and many users are able to identify unnecessary data permissions, and convey to the AI agent to never share that information.

We note that the number of participants granting permissions to automatically share data decreases. Specifically, the number of participants who never express *always share* permissions increases from 17.2% (when participants are presented with necessary data) to 25.6% (when participants are presented with unnecessary data, along with necessary data). When participants do express automatic sharing/non-sharing permissions, we do not observe substantial changes in their *always share* preferences. We surmise that when some users observe agent mistakes, they are more cautious in expressing preferences for automatic data sharing.

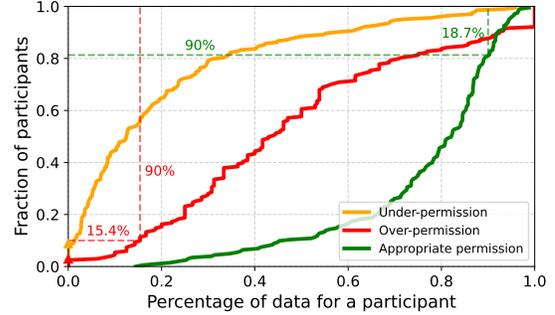


Figure 2: Distribution of *under-*, *over-*, and *appropriate* data sharing permission rate of each participant.

4.2.2. We observe that over-permissioning is substantially more common than under-permissioning. We also note that while participants struggle with providing appropriate permissions and often over-share, they are mostly cautious about sharing sensitive data. Under- and over-permissioning have been persistent problems in prior systems, such as mobile platforms, where users often granted too many or too few permissions due to poor understanding, misleading interfaces, or decision fatigue [22], [37], leading to privacy risks and/or broken functionality [3], [77]. Motivated by these challenges, we next examine whether similar issues arise in the context of AI agents, where the opacity of AI agents’ execution may make permission alignment even more challenging.

Figure 2 presents the distribution of under-, over-, and appropriate data sharing permission rates of each user.⁶ We find that only 8.9% of participants never engage in under-permissioning and 3.0% of participants never engage in over-permissioning. We observe that over-permissioning is substantially more common than under-permissioning, where 90% of participants engaging in over-permissioning for 15.4% or more data sharing requests they encounter. None of the participants give appropriate permissions for all of the data types on which they were prompted, and only 18.7% of the users give appropriate permissions for 90% or more data types on which they were prompted.

Through more in-depth analysis, we find that participants’ data sharing behavior is impacted by the sensitivity of data. For example, highly sensitive information such as *SSN* exhibits an under-permission ratio of about 46.3%, meaning participants are cautious with sharing such data. In contrast, for relatively less sensitive data, such as *Meeting Details*, which has an appropriate permission of 98.4%, participants are much more comfortable granting access. For ambiguous but non-sensitive data types, such as *Workspace Name* (generally used by Slack), participants tend to over-permission for 88.9% of the instances, likely because they may be uncertain about the actual need of this information.

6. *Appropriate permission* considers participant that provided only the data that was necessary (as determined during query curation§ 4.1.2) for solving the query, *over-permission* considers participants that shared data that was not required to resolve the query, and *under-permission* considers participant that withheld data that was necessary to address the query.

Domain	All		Concerning domain	
	Always	Never	Always	Never
Entertainment	55.6%	2.9%	60.5%	2.6%
Health&Fitn.	48.9%	7.7%	44.6%	9.7%
Smart Home	41.6%	9.3%	43.5%	7.6%
Travel	36.0%	10.5%	20.3%	3.8%
Shopping	32.2%	12.0%	24.0%	9.3%
Work&Produc.	30.8%	8.6%	27.6%	7.9%
Social	30.4%	8.1%	33.6%	9.9%
Finance	22.2%	16.9%	22.4%	13.9%
All	33.8%	10.7%	29.2%	10.8%

TABLE 1: Permission preferences (avg.) for always and never share permissions. *Concerning domain* columns contain data only from participants who mark the corresponding domain as concerning.

Additionally, we analyze some of the most frequently under-permissioned data types and find that participants are less likely to share sensitive personal information. For example, *Child Name* was not shared 65.1% of the time, *Work Information* by 57.4%, and *Passport Information* by 54.0% across all data permission requests. These data types are often essential for tasks, such as identity verification or financial transactions. Importantly, participants who did not share these data types frequently reported strong privacy concerns in relevant domains. For instance, 50.0% (15) of participants who did not share *Account Password* and 36.8% (25) who did not share *SSN* listed *Finance* as one of their privacy concerning domains. This suggests that participants are making intentional privacy-conscious choices, particularly in areas they value most. While some of these data types, such as *SSN* or *Passport Information*, are often essential for providing functionality, we find that others, such as *Child Name* and *Travel Details*, could be reasonably substituted with dummy values in some scenarios to maintain functionalities while respecting users’ privacy.

4.2.3. Participants’ permission preferences vary even within various contexts/domains, thus requiring finer contextual considerations. Participants are also able to identify and correct AI agents’ mistakes in sharing data. To make automatic decisions on users’ behalf, an AI agent needs to understand the factors that influence users’ permission decisions. We dive into several factors.

Communication Context/Domain. Significant prior research has identified the “context” of the communication as a key factor in user decision-making around data sharing [2], [8], [10], [36], [56]. Table 1 presents the eight contexts (which we refer to as domains) for which we record user preferences. We observe considerable variation in participants’ permission preferences in *always share* and *never share* permissions across different domains. For data sharing, we note that participants express the most *always share* permissions (i.e., 55.6%) for the *Entertainment* domain and the least permissions (i.e., 22.2%) for the *Finance* domain. For the *never share* permissions, the inverse is true, i.e., participants express their preferences to not share data

Domain	Tool	Always	Yes (Once)	No (Once)	Never
Entertainment	Movie Database	58.3%	33.7%	5.9%	2.1%
	Web	34.4%	52.5%	4.9%	8.2%
	Apple Music	68.3%	30.2%	1.6%	0.0%
Travel	Weather	50.8%	37.9%	5.1%	6.1%
	Local Search	42.9%	39.0%	9.9%	8.2%
	Travel Booking	32.6%	45.4%	9.6%	12.5%
	Calendar	43.9%	43.9%	5.3%	6.9%
	Cloud Drive	17.8%	56.3%	9.6%	16.4%
	Email	34.4%	50.8%	9.0%	5.7%
Work&Produc.	Slack	34.9%	47.6%	7.9%	9.5%
	Email	27.7%	53.6%	8.3%	10.4%
	Calendar	43.8%	47.4%	5.6%	3.2%
	Cloud Drive	22.6%	64.5%	7.3%	5.6%
Health&Fitn.	Fitness Tracking	48.7%	35.4%	9.0%	6.9%
	PregnancyPal	41.0%	42.6%	6.0%	10.4%
	Nutrition&Diet	46.2%	37.3%	7.5%	9.0%
	Hospital Booking	54.0%	33.9%	6.5%	5.6%
	Calendar	54.8%	35.5%	4.8%	4.8%
Social	SMS	29.2%	51.4%	7.6%	11.7%
	Cloud Drive	27.9%	54.1%	13.1%	4.9%
	Instagram	26.2%	52.5%	14.8%	6.6%
	Tinder	32.2%	50.0%	10.0%	7.7%
Finance	Banking	24.8%	50.6%	11.3%	13.2%
	Tax Management	22.6%	50.2%	9.2%	18.0%
Shopping	Amazon	32.2%	45.0%	10.7%	12.0%
Smart Home	Lighting Control	54.8%	32.3%	3.2%	9.7%

TABLE 2: Permission preferences across domains and tools.

for only 2.9% of cases within the *Entertainment* domain, but in 16.9% of cases with the *Finance* domain. We also observe that participants tend to give fewer *always share* permissions, for domains that they self-report as privacy concerning. This variation is particularly noticeable for *Travel* and *Shopping* domains, where *always share* decisions drop by 15.7% and 8.2%, respectively.

Differences Within Communication Contexts/Domains.

Next, we analyze the variance in participants’ permission preferences within domains. We present the breakdown of all participants’ permission preferences for data-sharing for different tools within different domains in Table 2. Overall, we note that participant preferences on data sharing for tools have differences within domains. Barring domains with single tools, *Entertainment* and *Travel* have the highest standard deviation of 14.2% and 10.5% for *always share*, and *Finance* has the lowest standard deviation of 1.1%. In the case of *Entertainment*, *Web* browsing tool gets the least *always share* permission preferences, and in the case of *Travel*, *Cloud Drive* gets the least *always share* permission preferences. Our investigation reveals that, in the context of *Travel*, *Cloud Drive* is always used to access users’ travel documents (e.g., passport scans, visa records). While participants allow access to these documents as they are essential for some travel tasks, they do not prefer AI agents to automatically retrieve and share such sensitive data.

For one-time sharing, *Entertainment* and *Finance* have the highest and lowest standard deviation of 9.8% and 0.2%, respectively. For never sharing and one-time non sharing permission preferences, the standard deviation between tools across domains remains mostly stable, and does not exceed 3.9% for never sharing and 2.8% for one-time non sharing.

These results indicate that permission preferences can

GT	PA	Always	Yes (Once)	No (Once)	Never
Necessary	Necessary	41.8%	47.1%	4.9%	6.1%
Necessary	Unnecessary	31.1%	36.0%	15.2%	17.7%
Unnecessary	Necessary	30.2%	48.2%	12.0%	9.5%
Unnecessary	Unnecessary	16.6%	21.0%	21.1%	41.4%

TABLE 3: Participant permission prefs. for perceived necessary/unnecessary data (PA), along with ground truth (GT).

vary even within domains, and thus granular context consideration is necessary for permission inferences, i.e., considering coarse context categories for understanding permission preferences across users may be insufficient. Our permission prediction model (in Section 5.1.3) thus encodes contextual information at a finer granularity.

Differences When Users Align and Misalign With the AI Agent. As we ask participants in our user study to select the *necessary data* that the AI agent will need to address a query, we develop a notion of participants developing an *understanding* of the AI agents’ execution. This setup allows us to analyze how user permission preferences may vary when they have some understanding of the AI agent’s execution. Table 3 presents participants’ permission preferences for perceived necessary and unnecessary data, along with the ground truth (as determined in Section 4.1.2). We observe that when participant perceptions of necessary and unnecessary data align with the ground truth (1st and 4th row in Table 3), the rate of appropriate permission decisions (allowing necessary data and denying unnecessary data) is high. In such cases, participants are more likely to express appropriate permissions, with an 88.9% and 62.5% correct decision rate for sharing and not sharing. We also note that, even when participants believe certain necessary data is unnecessary (2nd row in Table 3), they are still able to provide appropriate permissions, likely because data selection for specific queries provides more context for making an informed decision. Conversely, when the agent makes incorrect decisions, i.e., presents unnecessary data as necessary (3rd row in Table 3), 78.2% of times users share data with the AI agent, which implies that many users tend to believe the AI agent even when they are incorrect.

4.2.4. Participant demographics information (e.g., age) and self-reported metrics (e.g., privacy consciousness) correlate with their permission decision preferences.

Next, we analyze participants’ demographic information (i.e., age, education, and sex) and self-reported metrics (i.e., AI familiarity, AI usage frequency, AI trust, and privacy consciousness) to analyze how these factors influence participants’ permission preferences. As for demographics, we observe that the tendency to give *Always Allow* data sharing permissions decreases with participant age, correlating with generational differences in privacy attitudes as also observed by prior work [2]. We observe that there is no substantial difference in the permission preferences of male and female participants. As for self-reported metrics, higher AI familiarity and usage correlate with increased data-sharing preferences. We provide a more detailed analysis of demographics and self-reported metrics in Appendix A.

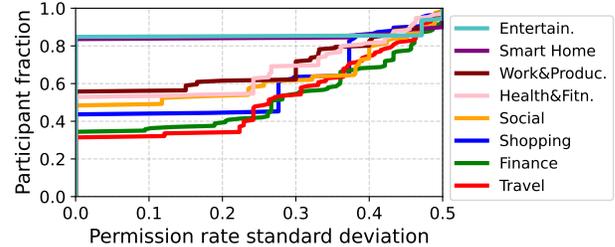


Figure 3: Distribution of standard deviation (SD) in participant permission preferences within domains. SD of 0 implies that participants choose the same preference within a domain, 0.1 implies there are 2 or 3 data types that do not align with other data types within a domain.

4.2.5. Individual participants’ permission preferences within a domain are often consistent; however, for some privacy-conscious participants, there is a high variance in their permission decisions. For an AI assistant to learn users’ permission preferences and apply them in various situations, an important criterion is that users’ decision-making is predictable and consistent. To that end, we analyze the consistency in participants’ permission decisions. From Figure 3, we note that within a domain, participants’ permission decisions are often consistent. The standard deviation is lowest for the *Entertainment* and *Smart Home* domains and highest for the *Travel*, *Finance*, and *Shopping* domains. Notably, many participants have entirely consistent data-sharing preferences within specific domains. There are 84.5%-85.6% of participants who are willing to either allow or deny all data permission for the *Entertainment* and *Smart Home* domains, 49.2%-56.4% for the *Work & Productivity*, *Health & Fitness*, and *Social* Domains, and 32.1%-44.4% for the *Shopping*, *Finance*, and *Travel* domains.

Additionally, we notice that some participants have relatively high intra-domain preference variance. We therefore analyze the top 10 participants with high intra-domain preference standard deviation and find that the wide variation for these participants can be traced to several key personal attributes, particularly AI trust levels, privacy concerns, and AI familiarity. Specifically, 8 participants report low to moderate trust in AI, even though 7 participants use AI frequently, and all rate privacy as very important. For these participants, the sensitivity of data types seems to matter when they make permission decisions. For example, within the *Finance* domain, participants often deny access to highly sensitive personal data such as SSNs, bank account numbers, and online account credentials, while allowing access to relatively less sensitive data, such as tax filing status, savings goals, or payment notes. Such a distinction leads to varied preferences for participants even within a single domain.

4.2.6. Participants’ permission preferences tend to follow similar patterns, presenting an opportunity to predict individual user preferences by leveraging insights from other users. In a real-world deployment, we anticipate that users will encounter a range of scenarios that may not

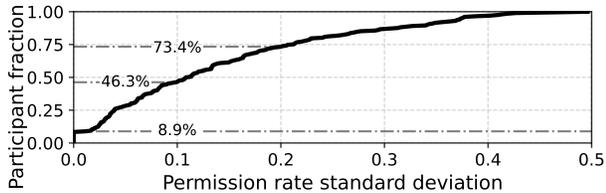


Figure 4: Distribution of SD in participant permission preferences across domains. SD of 0.1 means that preferences are mostly consistent, and SD of 0.2 means that only 1 or 2 domains have a noticeable difference with others.

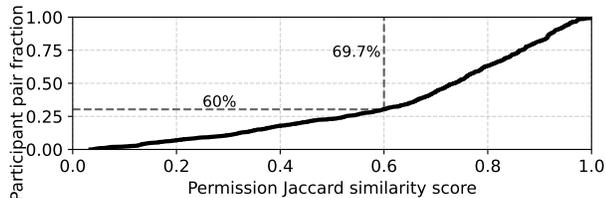


Figure 5: Jaccard similarity for participant pairs within groups. We group participants by the query set they answer.

arise during the training of a personal permission assistant. For example, to not fatigue the users, a deployed permission assistant may record a user’s preferences in one, two, or a handful of domains, but users’ day to day questions may span more domains. Thus, we seek to understand how transferable users’ permission decision making patterns are across contexts. Figure 4 presents the standard deviation of permission allowance rate across domains for users (Figure 9 in Appendix B shows the distribution of average permission allowance rates across domains for users.). We note that 46.3% of participants have a standard deviation below 0.1, which means their permission is mostly consistent for all the tested domains; 73.4% of participants have a standard deviation below 0.2, which means only one or two domains have a noticeable difference with other domains. Moreover, we find that 8.9% of participants maintain consistent permission preferences across all tested domains. We also find that over 35.5% of participants (not explicitly represented in Figure 4) show zero preference variance in at least half of the domains they interact with, suggesting that they tend to grant similar permissions across different contexts.

We also explore the consistency in permission preferences for data types across participants, as it can allow us to learn preferences from groups of users with similar preferences and apply them to other users. Figure 5 presents the Jaccard similarity between participants’ permission preferences on the same data permission requests. We note that participants’ permission preferences for the same permission requests are not unique and in fact, resemble many other users. For example, 69.7% of participant pairs (two participants answering the same permission requests) have similar data sharing preferences for 60% or more data types. To further understand user variance and consistency in data permission preferences, we analyze which data types exhibit

Top 10 high std. dev. data types		Top 10 low std. dev. data types	
Data type	Std. dev.	Data type	Std. dev.
Account Credentials	0.498	Music Listening History	0.125
Employment Details	0.498	Hobbies and Interests	0.178
Driver License Number	0.489	Relationship Preferences	0.212
Travel Itinerary	0.486	Test/Diagnostic Results	0.244
Passport Documents	0.485	Fitness Goal	0.244
Investment Information	0.483	Movie Preferences	0.246
Payment Method Details	0.482	Personal Biography	0.248
Bank Account Details	0.478	Travel Preferences	0.251
Social Security Number	0.475	Accommodation Details	0.255
Sender Email Address	0.461	Travel Destination	0.259

TABLE 4: Data types with high and low standard deviation in participants’ data permission preferences.

the highest and lowest variability across participants. As shown in Table 4, high-variance types like *Account Credentials*, *Employment Details*, and *Driver License Number* (with standard deviations of 0.49) indicate strong disagreement—likely due to the sensitivity of data. In contrast, low-variance types such as *Music Listening History* and *Hobbies and Interests* (with standard deviations as low as 0.125) reflect broad agreement, possibly because they are seen as low-risk and frequently shared in everyday contexts. These patterns suggest that preferences for some data types are relatively predictable, while others show diverse responses and are more challenging to predict.

5. Predicting User Permission Preferences

In this section, we explore the feasibility of predicting users’ permission decisions by leveraging data from our user study. We explore the potential of using *individual* user data as well as leveraging data from *other users* in developing an accurate permission prediction model. We also explore the role of various factors, such as the domains/contexts, individual data types, and variance in user permission preferences, in the accuracy of permission prediction.

We anticipate a sustained progression in permission modeling for AI agents over the next several years, during which a wide variety of models will be explored. In this paper, we focus more on the implications from our user study results and present our *permission assistant* as a proof of concept and/or initial exploration of the design space, rather than a definitive solution.

5.1. Designing a Permission Prediction Model

Our findings in Section 4.2 indicate that although various factors influence participants’ permission decisions, these influences ultimately result in relatively consistent participant permission decisions. Thus, we investigate whether users’ prior permission decision history, their demographics, and other self-reported attributes can be used to predict their future permissions decisions.

A key challenge is that we possess limited data on users, which may make it challenging to train a classifier that attains a high accuracy. To this end, we explore a hybrid

machine learning framework that relies on LLM-based in-context learning [57] and collaborative filtering [29]. We rely on LLM-based in-context learning because it can attain a high accuracy even with a handful of examples (i.e., “few-shot”) [19], [57]. We rely on collaborative filtering because it allows us to learn from other users with similar permission decision history [29], [30], thus complementing the limited permission history we possess on individual users.

5.1.1. In-context learning. As we observe in Section 4.2.5 and 4.2.6, participants’ permission decisions remain consistent within a domain and are transferable across domains, we first attempt to learn preferences only from users’ own permission decisions. To that end, we design an LLM-based in-context learning framework using OpenAI’s o3-mini reasoning-based LLM. To condition the LLM, we rely on role-based prompting, user demographics, other self-reported information, and user permission history.

Since our goal is to design an assistant that will assist users in their permission decision making, we adopt *permission assistant* as a role for the LLM. As for the demographics we include, users’ age range, education, and gender. For self-reported information, we mainly consider users’ AI familiarity (e.g., usage frequency, usage purposes) and privacy consciousness (e.g., value to privacy, trust in AI). User permission history includes the query, data type, and the name of the tool or AI agent requesting the data, along with the user’s permission decision. As LLMs take natural language input, we encode these features as natural language instructions. For example, demographic information is encoded as follows: *a male in the 45–54 age group with a bachelor’s degree*. As an output, we condition the LLM to provide a prediction label along with a confidence score, which prior research has shown to be reliable [62].

5.1.2. Collaborative filtering. Motivated by our observation in Section 4.2.6, where groups of users exhibit similar permission-granting behaviors across domains and scenarios, we explore collaborative filtering in predicting user preferences. Collaborative filtering is widely adopted in recommendation systems due to its effectiveness in capturing implicit user-user similarities based on commonalities in prior user preferences [29], [30]. We model user preferences as an adjacency matrix, where columns represent data types across contexts, rows represent users, and values in each cell represent either positive (sharing) or negative (non-sharing) permission preference. Since users do not provide their preferences on all data types, several of the cell values are empty, and our classification task is to predict the values of the cells. We rely on model-based collaborative filtering, which uses a light graph convolution network (LightGCN) [29].

LightGCN constructs a graph where users and permission requests (with context) are nodes, and edges represent observed permission preferences. It learns embeddings for users and permission requests by iteratively averaging information from their connected neighbors. The final prediction score is computed using the dot product between a user and a permission request embedding, indicating the likelihood

that the user would grant the permission. This score is then converted into a predicted label (i.e., allow or deny) by applying a threshold. The prediction threshold is configurable for collaborative filtering, which we currently set to equalize FPR and FNR (Appendix C provides more details).

5.1.3. Hybrid model using in-context learning and collaborative filtering. We next explore combining the in-context learning and collaborative filtering models to incorporate learning from both: (1) individual user preferences, which in-context learning incorporates, and (2) similar user preferences, which collaborative filtering incorporates. We combine these models by extending our in-context learning framework (described in Section 5.1.1). Specifically, we include the results from a collaborative filtering model as textual descriptions for the LLM. For example, a recommended permission decision is represented as follows: *<Query: Can you retrieve my tax filing details from last year?; Tool: Tax Management; Data Type: SSN; Decision: Deny>*. Since preferences for different participants can vary within domains (as shown in Section 4.2.3), such representation at the query level (i.e., finer granularity) allows us to naturally capture consistent preferences for individual participants.

While integrating collaborative filtering results, we only consider high-confidence permission predictions. We consider high-confidence predictions to be the ones where both the false positive rate (FPR) and false negative rate (FNR) are below 5%, which covers 35.0% of the permission requests. Note that in our testing, we tried including all CF predictions (i.e., both high- and low-confidence predictions) in our hybrid model, but that deteriorates model accuracy.

Note that our collaborative filtering alone is incapable of making predictions on previously unseen data, and our in-context learning model only used data from individual users. Our hybrid model allows collaborative filtering and in-context learning to complement each other. As a result, it yields a personalized predictor for each user (i.e., as many personalized instances as users), composed of shared user-specific CF parameters as well as an in-context learning model trained on individual user data.

5.2. Evaluation

5.2.1. Datasets and metrics. We construct our training and testing datasets using responses collected from the user study to evaluate the effectiveness of our permission prediction model. We only consider decisions marked as *yes, always share* or *no, never share* as ground truth for sharing and non-sharing data. Additionally, we exclude participants who specified always/never sharing preferences for fewer than five scenarios/queries. After filtering, our dataset contains 7,563 permission decisions from 181 participants, including 5,244 allowance and 2,319 denial responses. We define a positive outcome as permission being granted and a negative outcome as permission being denied. For model training and evaluation, we perform 5-fold cross-validation. Specifically, the user-answered questions are split into five folds; in each iteration, we train on four folds and test on

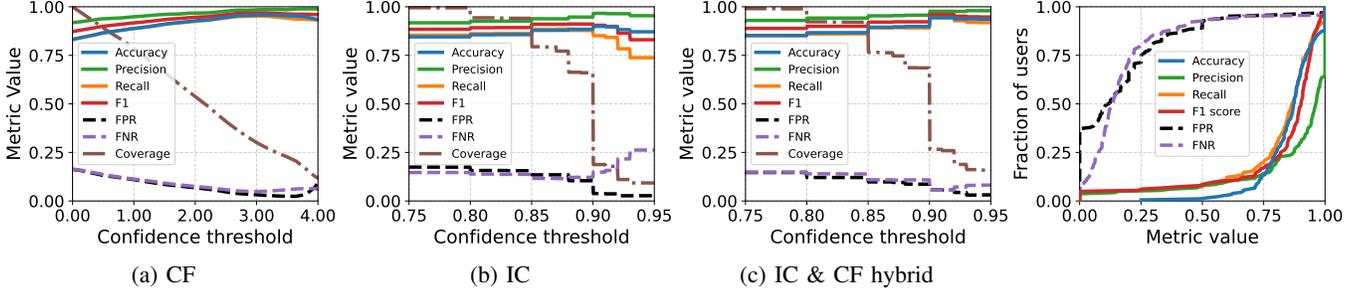


Figure 6: Distribution of classification metrics. For IC & hybrid models, the confidence score is reported by the LLM. For CF, the confidence score is the difference b/w prediction thresholds that reduce FPR and FNR, with larger values implying higher confidence. Figure 7: Distribution of hybrid model metrics over the fraction of users.

Metric	CF (%)	IC (%)	IC & CF (%)
Accuracy	83.3±1.4	84.4±0.7	85.1±0.4
Precision	91.9±0.5	91.5±1.1	92.8±0.7
Recall	83.3±1.4	85.4±1.5	85.2±1.2
F1 score	87.4±1.0	88.3±0.7	88.8±0.5
FPR	16.7±1.5	17.9±2.3	15.2±2.1
FNR	16.7±1.4	14.6±1.5	14.8±1.2

TABLE 5: Classification metrics with full coverage across different model configurations.

the remaining fold. This is repeated five times, so every question is tested exactly once and is never used for training and testing in the same iteration. This setup allows us to evaluate the models on unseen data.

5.2.2. Classification confidence threshold. We start by comparing model configurations: (1) a model trained on individual user data using in-context (IC) learning, (2) a collaborative filtering (CF) model trained on the entire permission-granting history of all users, and (3) a hybrid of in-context learning and collaborative filtering. Figure 6 presents the distribution of classification metrics for various model configurations. Overall, we attain the highest accuracy (95.8%) for CF with a confidence threshold of 3.17. However, we compromise on *coverage*⁷ as we are only able to make predictions on 27.3% of the data. For IC, we attain an accuracy of 89.9% with a confidence threshold of 0.90, and a coverage of only 18.8%. For the hybrid model of IC and CF, we attain an accuracy of 94.4% with a confidence threshold of 0.91, and a coverage of 25.9%. These trends indicate that, as the confidence threshold gets stricter, it results in precise predictions but reduces the coverage.

Classification Accuracy. To make classification decisions for all data, i.e., for 100% coverage, across all model configurations, we are able to achieve an accuracy of 83.3%, 84.4%, and 85.1%, for CF, IC, and the hybrid IC and CF model. Table 5 presents other classification metrics for full data coverage. We note that the hybrid model outperforms

7. We define *coverage* as the fraction of permission requests for which the model can recommend an action with confidence above a specified threshold. Lower coverage, however, also means that users must make more decisions manually, and only high-confidence cases are automated.

Hist. ratio	0%	25%	50%	75%	100%
#Queries	0	1-4	2-8	3-12	4-16
Accuracy (%)	66.9±1.6	77.7±1.3	82.1±0.7	83.7±0.8	85.1±0.4
Precision (%)	83.0±1.4	87.6±1.2	90.0±0.9	91.1±1.0	92.8±0.7
Recall (%)	65.7±0.7	79.1±1.5	83.6±0.6	84.7±0.7	85.2±1.2
F1 score (%)	73.4±0.9	83.1±1.0	86.6±0.4	87.8±0.7	88.8±0.5
FPR (%)	30.6±3.8	25.4±3.2	21.2±2.8	18.7±2.5	15.2±2.1
FNR (%)	34.3±0.7	20.9±1.5	16.4±0.6	15.3±0.7	14.8±1.2

TABLE 6: Impact of permission history on classification.

individual model configurations on all metrics, except for recall and FNR. Notably, the FPR decreases by 2.7% when recommendations from the CF model are taken into account. These results indicate that incorporating contextual examples from the CF model enhances prediction accuracy. However, as we note in Section 5.1.3 that these gains strongly correlate with the quality of CF predictions.

As permission assistants will make predictions to share data on behalf of users, a high precision and lower FPR may be desired, which requires a compromise on model coverage. Thus, in practice, the users may need to be involved in the loop for making data permission decisions (elaborated in Section 7). Moreover, as we find (from Section 4.2.5) that users express varying privacy concerns across domains, custom confidence thresholds could be configured at the granularity of individual users and/or domains. We also find (from Section 4.2.6) that certain data types show a high degree of variance across users, so they may also be excluded from predictions to avoid mistakes.

5.2.3. Impact of permission decision history. In a real-world setting, users may not be expected to provide permission preferences for a large number of data types. Thus, we explore the relation between number of training samples available per user and accuracy of our permission assistant.

To assess the impact of permission decision history on prediction accuracy, we incorporate different ratios of the permission decision history into each user’s model context and evaluate the corresponding performance. Table 6 shows the metrics for our hybrid model under various ratios of permission decision history. We note that even without any permission history, the model still infers users’ preferences

by leveraging users’ demographic information, AI usage experience, and privacy concerns. We also observe that the transition from no permission history to even a small amount (i.e., permission history on 1–4 queries) yields significant improvements (i.e., more than 10.8% increase in accuracy).

Overall, our results indicate that an increase in permission decision history (i.e., an increase in training data) results in improved prediction accuracy. This continuous learning capability can facilitate real-world deployment, i.e., as users naturally generate more permission history over time, the model can iteratively learn from them, leading to increasingly accurate permission predictions.

5.3. Result Analysis

5.3.1. Per-user performance analysis. As each user has their personalized model for predicting permission decisions, we analyze how these models perform on permission decision predictions for each user. Specifically, we compute per-user metrics using the prediction results from each user’s dedicated hybrid IC and CF model, and attempt to identify the factors that contribute to lower or higher accuracy.

Overall Trends. Figure 7 presents the distribution for performance metrics computed over individual users. Notably, 35.4% of users achieve accuracy greater than 90%, and 12.7% of users even achieve perfect accuracy, meaning every permission decision predicted by their dedicated model is correct. For precision, 70.4% of users reach values greater than or equal to 90%, suggesting that these models are reliable in correctly predicting data sharing (i.e., predicting positive cases). In contrast, 34.8% of users achieve recall values of 90% or higher, indicating that models are generally conservative in making automated data sharing predictions.

The error metrics further elaborate on model performance. The distribution for FPR shows that 48.2% of users’ false permission grants (i.e., FPR) are at or below 10% (with 0 false permission grants for 37.2% of users). Meanwhile, 34.8% of users have an FNR of 10% or less, indicating that more than a third of the users’ models are able to constrain the number of false permission denials. Overall, the personalized models show promising performance in predicting individual users’ permission decisions. In the remainder of this section, we analyze the errors to gather insights for further improving the models.

Detailed Analysis. We perform an in-depth analysis of the bottom 20% (36) users with the lowest accuracy by comparing this group to the top 20% of users with the highest accuracy. There are significant differences in the average performance metrics between these two groups, for instance, average accuracy (97.6% vs. 69.7%), recall (98.2% vs. 62.3%), and precision (98.0% vs. 69.2%). Next, we investigate the factors that may contribute to these differences. We first examine whether users’ self-reported attributes show significant discrepancies. While some trends are observable, for example, high-accuracy users report slightly lower AI trust levels (2.39 vs. 2.81 on average), higher AI usage frequency (3.19 vs. 3.08), and lower privacy concerns (4.06

vs. 4.14), these differences are not statistically significant (Mann-Whitney U test p -values range from 0.07 to 0.95). We then examine whether there are significant differences in the number of permission decision history entries and the number of collaborative filtering recommendations (as we observe them to be key factors in improving prediction accuracy in Section 5.2.2 and 5.2.3). We observe notable differences in the average number of permission history queries between the two groups (11.69 for high-accuracy users vs. 9.38 for low-accuracy users, U-test $p = 0.01$), as well as in the number of CF recommendations (3.96 vs. 2.90, U-test $p = 0.07$).

Next, we analyze the standard deviation in permission allowance rates across domains for both high-accuracy and low-accuracy users. We note that high-accuracy users tend to have lower variability in their permission preferences across domains, with a larger proportion of users exhibiting smaller standard deviation values, indicating more consistent preferences. In contrast, low-accuracy users show greater variability in their permission preferences across domains. For instance, 30.6% of high-accuracy users fall at the lower end of the standard deviation range (≤ 0.04), compared to just 11.1% of low-accuracy users. Similarly, at the upper end (≥ 0.27), 22.2% of low-accuracy users exceed this standard deviation threshold (0.27), while only 13.9% of high-accuracy users do. We observe that high standard deviation in users’ permission decisions may make it more challenging to accurately predict their permission preferences. For instance, predicting the preferences of a user with a standard deviation of 0.43 yields subpar performance, with an accuracy of 49.0%.

A closer look at this user’s prediction results reveals the impact of such variance: although the user is generally willing to share data in domains like *Smart Home*, *Travel*, and *Health & Fitness*, they tend to deny data sharing in *Social*, *Finance*, and *Shopping*. We surmise that when the model sees more permission history from the first group of domains, it tends to infer a higher allowance rate even for the latter group, and vice versa, during five-fold cross-validation. This suggests that for users with high variance in their permission preferences, patterns learned from some domains may not transfer well to others, resulting in reduced prediction performance. In contrast, when users have more consistent permission preferences across domains (i.e., lower variance), the model finds it easier to transfer learned patterns. For example, a high-accuracy user with a standard deviation of just 0.01 achieves an accuracy of 98.0%, recall of 100.0%, and precision of 97.8%. We note that it is a key challenge to make predictions of users’ permission preferences when there is a high degree of variance in their prior permission decisions.

Finally, we investigate the impact of CF recommendation accuracy on the two user groups. We find that high-accuracy users have CF recommendations with an accuracy of 99.5%, compared to 89.5% CF recommendation accuracy for low-accuracy users. Additionally, when CF recommendations are correct, the model consistently follows them, showing 100% alignment for low-accuracy users and 99.8% for high-

accuracy users. However, when CF recommendations are incorrect, the model still tends to follow them. This highlights the importance of filtering out unreliable CF results to avoid misleading the model’s final decisions. For CF predictions to improve, it is crucial that users can be grouped with others who share similar preferences. Thus, as more data is collected—particularly from a larger and more diverse user base—the quality of CF predictions naturally improves.

5.3.2. Contextual performance analysis. Next, we analyze how well the model predicts user preferences under different contexts, specifically across domains, tools, and data types. Our goal is to examine whether there are significant differences in accuracy, FPR, and FNR among these contextual factors, thereby revealing patterns in when and where the model tends to be over- or under-permissive.

Performance by Domain. When analyzing performance across different domains, we find that the model performs slightly better at predicting permission requests in certain domains. Notably, *Work & Productivity* (87.4% accuracy), *Health & Fitness* (87.1%), and *Entertainment* (86.8%) exhibit relatively high accuracies. In contrast, domains such as *Shopping* (84.3% accuracy), *Finance* (81.5%), and *Smart Home* (80.1%) show lower prediction accuracies.

We find that these results correlate with the variance of permission decisions in domains (Figure 3 plots standard deviation across domains). Specifically, the high-accuracy domains have lower variance in participants’ permission preferences and low-accuracy domains have the highest variance in participants’ permission preferences. A closer examination of the *Smart Home* results reveals a high FNR of 23.8%, indicating a conservative bias in which the model tends to reject permissions that users might have accepted in this domain. Notably, *Finance* has the highest FNR of all domains, at 28.3%, which reflects the model’s cautious behavior in this particularly sensitive domain. By contrast, the *Social* domain exhibits the highest FPR, at 23.3%, suggesting that the model is more likely to approve permission requests in less sensitive domains.

Overall, we note that while variance within domains degrades accuracy, the degradation is not substantial, especially as compared to the degradation in accuracy with variance within the user’s permission decisions (Section 5.3.1).

Performance by Tool. At the tool level, disparities become even more apparent. Tools with the highest accuracies include *Calendar* (92.0%), *The Movie Database* (89.0%), *Tinder* (88.8%), *Hospital Booking* (88.6%), and *Weather* (88.4%). These are also the tools for which participants tend to grant more permissions. For example, as shown in Table 2, *Calendar* appears in multiple domains such as *Travel*, *Work & Productivity*, and *Health & Fitness*, and it ranks among the top tools with the most *Always allow* permissions. A similar pattern holds for the other aforementioned tools, suggesting that the model predicts permissions more accurately when there is more data available, and users have consistent preferences for specific tools.

In contrast, some tools have noticeably lower accura-

cies, including *SMS* (76.4%), *Cloud Drive* (76.4%), *Slack* (79.3%), *Lighting Control* (80.0%), and *Banking* (80.3%). Among these tools, *Cloud Drive* (FNR of 31.4%), *Banking* (27.3%), and *Lighting Control* (20.9%) suffer from high FNRs, reflecting the model’s tendency toward caution.

Performance by Data Type. We observe substantial differences while analyzing the model’s accuracy across data types. For instance, some data types yield extremely high accuracies, even reaching 100%, such as *Fitness Goal* and *Hobbies and Interests*. Other data types with similarly strong results are *Relationship Preferences* (98.1%), *Personal Biography* (96.2%), *Travel Destination* (95.5%), and *Accommodation Details* (95.4%). It is worth noting that these data types also rank among those with the lowest preference variance among participants, as shown in Table 4. This supports our hypothesis that data types with more consistent user preferences are easier for the model to predict accurately.

On the other hand, data types like *Payment Method Detail* (59.1%), *Employment Details* (63.6%), *Driver License Number* (71.4%), and *Travel Itinerary* (75.9%) show some of the lowest accuracies. These data types also have the highest variance in user preferences, according to Table 4. In other words, data types with more diverse permission preferences tend to be more difficult for the model to predict. Other low-accuracy data types often involve highly sensitive information as well, such as *Personal Identification Numbers* (61.9%) and *Travel History* (63.2%). These types of data also tend to have higher FNRs. For instance, *Payment Method Details* has an FNR of 87.5%, *Personal Identification Numbers* has 72.2%, and *Driver License Number* has 66.7%. These are among the most under-permissioned data types. Only a few show relatively high FPRs, such as *Employment Details* and *Travel History*, both at 28.6%. These results suggest that the model often misclassifies these sensitive data types as unacceptable to share, even when users may permit it. This reflects challenges in accurately capturing context-sensitive boundaries for personally identifiable or other sensitive information.

6. Discussion

Reinforcing and Expanding Prior Knowledge. Our findings on user permissions for AI agents confirm established access control principles from traditional systems while also revealing agent-specific dynamics. Consistent with prior work on mobile applications and voice assistants [2], [43], [49], [63], [70], our results show that over-permissioning remains a key challenge, and that user decisions are highly dependent on context, such as data sensitivity and the information’s recipient. However, the autonomous nature of AI agents fundamentally alters the control dynamic. We find that permission granting becomes a continuous process of trust evaluation: when an agent makes a mistake, users are significantly less willing to grant permissions, indicating that agent performance is a new, critical factor influencing privilege. Most critically for our work, we find that despite these complexities, individual preferences exhibit a high

degree of consistency within specific contexts and across users. This predictable pattern is the key insight that enables our permission-prediction system, demonstrating that foundational privacy norms persist while the agent’s autonomy introduces new factors for permission management.

Model Robustness and Limitations. As agents interface with data and resources from potentially untrustworthy entities, it is crucial that the permission inference is robust against such attacks. When users communicate with AI agents using natural language instructions, they often reveal implicit or explicit preferences about what the agent is allowed to do. Such information can complement the execution context and past user behavior in making accurate predictions. In fact, prior work has used such information to predict the expected control and data flow of AI agents and detect anomalies [16], [79]. Similarly, prior work has proposed AI agent architectures that, by design, limit the flow of information between system modules [8], [79] and reliably control the generation of text in LLMs [52]. We believe that such approaches can be extended to support robust permission management in AI agents.

That said, natural language interactions can be ambiguous, and this remains an open problem for LLMs [45]. Our in-context learning model can be affected by this ambiguity and may make mistakes when processing unclear instructions. However, because we also rely on collaborative filtering over deterministic data types, the system inherently offers some resilience to such ambiguity for previously seen data types. Once a request is mapped to a canonical label, collaborative filtering predictions become insensitive to wording. Moreover, taking only high-confidence predictions and delegating uncertain data access permissions to users can help mitigate the model’s robustness issues. This approach pairs predictions with clear controls to make and revoke decisions and to provide feedback, which reduces interruptions for routine cases while allowing human oversight in riskier or uncertain situations.

Towards Usable Permission Management. Automated permission management is a multifaceted problem. Our work in this paper focuses on one facet, and other important facets, including improving the usability of permission management in AI agents, remain challenges and open avenues for future work. For example, during early use, the assistant may still need to learn a user’s preferences; a user’s preferences may change as their circumstances change (e.g., a previously trusted entity becomes untrusted); and some situations may be fundamentally difficult to predict. Thus, a full-fledged permission system will need to include not only a permission prediction module, but also UI/UX for engaging the user directly in cases where the predictive model has low confidence or makes a mistake—for example, UI/UX for users to make explicit permission decisions, to revoke previously-granted permissions, and to give feedback to the permission assistant. We believe there is rich future work to be done on how to design such a hybrid system well. The important contribution of our work here is to start this line of inquiry and to demonstrate that it is feasible: an

effective AI permission assistant can substantially reduce the number of decisions that users must be asked to make or evaluate directly, paving the way for a usable and secure permission management system.

7. Conclusion

In this paper, we explored the capabilities and limits of automating permission management in AI agents. Through a user study, we found that long-standing challenges such as over-permissioning persist in agentic systems, alongside new issues unique to them. Notably, when agents make mistakes, users become less willing to grant permissions, indicating that agent performance is a critical factor influencing privilege decisions. We also found that permission choices are strongly shaped by communication context, yet remain consistent within that context across user groups. Leveraging these insights, we developed a permission prediction model that combines collaborative filtering with LLM in-context learning, achieving 94.4% accuracy for high-confidence predictions. Even without prior permission history, the model reached 66.9% accuracy, with just 1–4 samples improving it by 10.8%. Our findings demonstrate the potential of learning user preferences to automate many permission decisions. However, challenges remain in enforcing predictions, improving robustness, and designing usable interfaces. Overall, this work advances automated permission management and outlines key directions toward its practical deployment.

Acknowledgment

We thank the reviewers for their valuable feedback. This work was partially supported by NSF (CNS-2154930, CNS-2238635), ARO (W911NF-24-1-0155), ONR (N000142412663), and gifts from Microsoft. T. Kohno was also supported by the McDevitt Chair in Computer Science, Ethics, and Society at Georgetown University; the majority of this research was conducted while he was at the University of Washington. We would also like to thank Pardis Emami-Naeini for providing feedback on the user study and statistical analysis.

References

- [1] Gpt action authentication. <https://platform.openai.com/docs/actions/authentication>, 2025. Accessed: 2025-06-02.
- [2] Noura Abdi, Xiao Zhan, Kopo M Ramokapane, and Jose Such. Privacy norms for smart home personal assistants. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.
- [3] Hazim Almuhiemedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your location has been shared 5,398 times! a field study on mobile app privacy nudging. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 787–796, 2015.
- [4] Marc Serramia Amoros, William Seymour, Natalia Criado, and Michael Luck. 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), 2023.

- [5] Android Developers. Permissions overview. <https://developer.android.com/guide/topics/permissions/overview>, 2023. Accessed: 2025-03-26.
- [6] Android Open Source Project. Runtime permissions. https://source.android.com/docs/core/permissions/runtime_perms, 2023. Accessed: 2025-03-26.
- [7] Anthropic. Developing a computer use model, 2024.
- [8] Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. Airgapagent: Protecting privacy-conscious conversational agents. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3868–3882, 2024.
- [9] Natã M Barbosa, Joon S Park, Yaxing Yao, and Yang Wang. “what if?” predicting individual users’ smart home privacy preferences and their changes. *Proceedings on Privacy Enhancing Technologies*, 2019.
- [10] Adam Barth, Anupam Datta, John C Mitchell, and Helen Nissenbaum. Privacy and contextual integrity: Framework and applications. In *2006 IEEE symposium on security and privacy (S&P’06)*, pages 15–pp. IEEE, 2006.
- [11] Igor Bilogrevic, Balazs Engedy, Judson L Porter III, Nina Taft, Kamila Hasanbega, Andrew Paseltiner, Hwi Kyoung Lee, Edward Jung, Meggyn Watkins, PJ McLachlan, et al. ”shhh... be {quiet!}” reducing the unwanted interruptions of notification permission prompts on chrome. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 769–784, 2021.
- [12] André Brandão, Ricardo Mendes, and João P Vilela. Prediction of mobile app privacy preferences with user profiles via federated learning. In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy*, pages 89–100, 2022.
- [13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [14] David Chan. So why ask me? are self-report data really that bad? In *Statistical and methodological myths and urban legends*, pages 329–356. Routledge, 2010.
- [15] Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. Personalized privacy assistants for the internet of things: Providing users with notice and choice. *IEEE Pervasive Computing*, 17(3):35–46, 2018.
- [16] Edoardo Debenedetti, Iliia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design. *arXiv preprint arXiv:2503.18813*, 2025.
- [17] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate attacks and defenses for llm agents. *arXiv preprint arXiv:2406.13352*, 2024.
- [18] Verena Distler, Matthias Fassel, Hana Habib, Katharina Krombholz, Gabriele Lenzini, Carine Lallemand, Vincent Koenig, and Lorrie Faith Cranor. Empirical research methods in usable privacy and security. In *Human Factors in Privacy Research*, pages 29–53. Springer International Publishing Cham, 2023.
- [19] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, 2024.
- [20] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 1388–1401, 2016.
- [21] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. Android permissions demystified. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 627–638, 2011.
- [22] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the eighth symposium on usable privacy and security*, pages 1–14, 2012.
- [23] Dorota Filipczuk, Tim Baarslag, Enrico H Gerding, and MC Schraefel. Automated privacy negotiations with preference uncertainty. *Autonomous Agents and Multi-Agent Systems*, 36(2):49, 2022.
- [24] Janet Finch. The vignette technique in survey research. *Sociology*, 21(1):105–114, 1987.
- [25] Google DeepMind. Introducing gemini 2.0: our new ai model for the agentic era, 2024.
- [26] Marian Harbach, Igor Bilogrevic, Enrico Bacis, Serena Chen, Ravjit Uppal, Andy Paicu, Elias Klim, Meggyn Watkins, and Balazs Engedy. Don’t interrupt me—a large-scale study of on-device permission prompt quieting in chrome. In *NDSS. The Internet Society, San Diego, CA*, 2024.
- [27] David Harborth and Sebastian Pape. Investigating privacy concerns related to mobile augmented reality apps—a vignette based online experiment. *Computers in Human Behavior*, 122:106833, 2021.
- [28] Harrison Chase and the LangChain Team. Langchain. <https://www.langchain.com/>, 2025.
- [29] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648, 2020.
- [30] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.
- [31] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
- [32] Umar Iqbal, Pouneh Nikkha Bahrami, Rahmadi Trimananda, Hao Cui, Alexander Gamero-Garrido, Daniel Dubois, David Choffnes, Athina Markopoulou, Franziska Roesner, and Zubair Shafiq. Tracking, profiling, and ad targeting in the alexa echo smart speaker ecosystem. In *ACM Internet Measurement Conference (IMC)*, 2023.
- [33] Umar Iqbal, Tadayoshi Kohno, and Franziska Roesner. Llm platform security: applying a systematic evaluation framework to openai’s chatgpt plugins. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 611–623, 2024.
- [34] Umar Iqbal, Tadayoshi Kohno, and Franziska Roesner. Llm platform security: applying a systematic evaluation framework to openai’s chatgpt plugins. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 611–623, 2024.
- [35] Jerry Liu and Simon Suo and the LlamaIndex Team. Llamaindex. <https://www.llamaindex.ai/>, 2025.
- [36] Yunhan Jack Jia, Qi Alfred Chen, Shiqi Wang, Amir Rahmati, Ear-lence Fernandes, Zhuoqing Morley Mao, Atul Prakash, and SJ University. Contextlot: Towards providing contextual integrity to affiliated iot platforms. In *ndss*, volume 2, pages 2–2. San Diego, 2017.
- [37] Patrick Gage Kelley, Sunny Consolvo, Lorrie Faith Cranor, Jaeyeon Jung, Norman Sadeh, and David Wetherall. A conundrum of permissions: installing applications on an android smartphone. In *Financial Cryptography and Data Security: FC 2012 Workshops, USEC and WECSR 2012, Kralendijk, Bonaire, March 2, 2012, Revised Selected Papers 16*, pages 68–79. Springer, 2012.

- [38] Leona Lassak, Hanna Püschel, Tobias Gostomzyk, and Markus Dürmuth. Introducing data trustees: A vignette-based study approach to get users in the loop. 2023.
- [39] Leona Lassak, Hanna Püschel, Oliver D Reithmaier, Tobias Gostomzyk, and Markus Dürmuth. Balancing privacy and data utilization: A comparative vignette study on user acceptance of data trustees in germany and the us. In *Network and Distributed System Security (NDSS) Symposium*, 2025.
- [40] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–31, 2018.
- [41] Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. Dissecting human and llm preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1790–1811, 2024.
- [42] Xiaotian Vivian Li, Mary Beth Rosson, and Jenay Robert. A scenario-based exploration of expected usefulness, privacy concerns, and adoption likelihood of learning analytics. In *Proceedings of the ninth ACM conference on learning@ scale*, pages 48–59, 2022.
- [43] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I Hong. Modeling {Users’} mobile app privacy preferences: Restoring usability in a sea of permission settings. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 199–212, 2014.
- [44] Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. We’re afraid language models aren’t modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, 2023.
- [45] Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. We’re afraid language models aren’t modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, 2023.
- [46] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Al-muhimedi, Shikun Aerin Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Twelfth symposium on usable privacy and security (SOUPS 2016)*, pages 27–41, 2016.
- [47] Gary Liu and Nathan Malkin. Effects of privacy permissions on user choices in voice assistant app stores. *Proceedings on Privacy Enhancing Technologies*, 2022.
- [48] Nathan Malkin, Serge Egelman, and David Wagner. Privacy controls for always-listening devices. In *Proceedings of the New Security Paradigms Workshop*, pages 78–91, 2019.
- [49] Nathan Malkin, David Wagner, and Serge Egelman. Runtime permissions for privacy in proactive intelligent assistants. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 633–651, 2022.
- [50] Manus AI. Manus - an autonomous artificial intelligence agent. <https://manus.im/>.
- [51] Ricardo Mendes, Mariana Cunha, João P Vilela, and Alastair R Beresford. Enhancing user privacy in mobile devices through prediction of privacy preferences. In *European Symposium on Research in Computer Security*, pages 153–172. Springer, 2022.
- [52] Michal Moskal, Madan Musuvathi, and Emre Kiciman. AI Controller Interface. <https://github.com/microsoft/aici/>, 2024.
- [53] Sara Motiee, Kirstie Hawkey, and Konstantin Beznosov. Do windows users follow the principle of least privilege? investigating user account control practices. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, pages 1–13, 2010.
- [54] Rajiv Movva, Pang Wei Koh, and Emma Pierson. Annotation alignment: Comparing llm and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9048–9062, 2024.
- [55] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. Privacy expectations and preferences in an {IoT} world. In *Thirteenth symposium on usable privacy and security (SOUPS 2017)*, pages 399–412, 2017.
- [56] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.
- [57] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.
- [58] OpenAI. Consequential flag. <https://platform.openai.com/docs/actions/production/consequential-flag#consequential-flag>, 2024.
- [59] OpenAI. Introducing gpts. <https://openai.com/blog/introducing-gpts>, 2024.
- [60] OpenAI. Chatgpt. <https://chatgpt.com/>, 2025.
- [61] OpenAI. Introducing operator, 2025.
- [62] Yudi Pawitan and Chris Holmes. Confidence in the reasoning of large language models. *Harvard Data Science Review*, 7(1), 2025.
- [63] Sai Teja Peddinti, Igor Bilogrevic, Nina Taft, Martin Pelikan, Úlfar Erlingsson, Pauline Anthonysamy, and Giles Hogben. Reducing permission requests in mobile apps. In *Proceedings of the internet measurement conference*, pages 259–266, 2019.
- [64] Prolific. Prolific - recruitment platform for online research. <https://www.prolific.com>.
- [65] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, Phillipa Gill, et al. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem. In *The 25th Annual Network and Distributed System Security Symposium (NDSS 2018)*, 2018.
- [66] Harry T Reis and Charles M Judd. *Handbook of research methods in social and personality psychology*. Cambridge University Press, 2000.
- [67] Talia Ringer, Dan Grossman, and Franziska Roesner. Audacious: User-driven access control with unmodified operating systems. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 204–216, 2016.
- [68] Franziska Roesner, Tadayoshi Kohno, Alexander Moshchuk, Bryan Parno, Helen J Wang, and Crispin Cowan. User-driven access control: Rethinking permission granting in modern operating systems. In *2012 IEEE Symposium on Security and Privacy*, pages 224–238. IEEE, 2012.
- [69] Jasmin Schwab, Alexander Nussbaum, Anastasia Sergeeva, Florian Alt, and Verena Distler. What makes phishing simulation campaigns (un) acceptable? a vignette experiment. In *Network and Distributed System Security Symposium, NDSS 2025*, 2025.
- [70] Suranga Seneviratne, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti. Your installed apps reveal your gender and more! *ACM SIGMOBILE Mobile Computing and Communications Review*, 18(3):55–61, 2015.
- [71] Yashothara Shanmugarasa, Hye-young Paik, Salil S Kanhere, and Liming Zhu. Automated privacy preferences for smart home data sharing using personal data stores. *IEEE Security & Privacy*, 20(1):12–22, 2021.
- [72] Han Shao, Xiang Li, and Guodi Wang. Are you tired? i am: Trying to understand privacy fatigue of social media users. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.
- [73] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. On the challenges in usable security lab studies: Lessons learned from replicating a study on ssl warnings. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, pages 1–18, 2011.

- [74] Rahmadi Trimananda, Hieu Le, Hao Cui, Janice Tran Ho, Anastasia Shuba, and Athina Markopoulou. {OVRseen}: Auditing network traffic and privacy policies in oculus {VR}. In *31st USENIX security symposium (USENIX security 22)*, pages 3789–3806, 2022.
- [75] Han Wang, An Zhang, Nguyen Duy Tai, Jun Sun, Tat-Seng Chua, et al. Ali-agent: Assessing llms’ alignment with human values via agent-based evaluation. *Advances in Neural Information Processing Systems*, 37:99040–99088, 2024.
- [76] Julia Wiesinger, Patrick Marlow, and Vladimir Vuskovic. Agents, 2025. <https://www.kaggle.com/whitepaper-agents>.
- [77] Primal Wijesekera, Arjun Baokar, Ashkan Hosseini, Serge Egelman, David Wagner, and Konstantin Beznosov. Android permissions re-mystified: A field study on contextual integrity. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 499–514, 2015.
- [78] Yuhao Wu, Evin Jaff, Ke Yang, Ning Zhang, and Umar Iqbal. An in-depth investigation of data collection in llm app ecosystems. In *ACM Internet Measurement Conference (IMC)*, 2025.
- [79] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. IsolateGPT: An Execution Isolation Architecture for LLM-Based Agentic Systems. In *Network and Distributed System Security (NDSS) Symposium*, 2025.
- [80] Jierui Xie, Bart Piet Knijnenburg, and Hongxia Jin. Location sharing privacy preference: analysis and personalized recommendation. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 189–198, 2014.
- [81] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.
- [82] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.
- [83] Nicole Zhan, Stefan Sarkadi, and Jose Such. Privacy-enhanced personal assistants based on dialogues and case similarity. In *European Conference on Artificial Intelligence*. IOS Press, 2023.
- [84] Xiao Zhan, Stefan Sarkadi, Natalia Criado, and Jose Such. A model for governing information sharing in smart assistants. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 845–855, 2022.
- [85] Yanjie Zhao, Xinyi Hou, Shenao Wang, and Haoyu Wang. Llm app store analysis: A vision and roadmap. *ACM Transactions on Software Engineering and Methodology*, 2024.

Appendix A. Analysis of Demographics and Self-Reported Metrics

We analyze each participant’s basic demographic information and self-reported metrics related to AI usage and privacy consciousness to investigate whether these factors significantly influence their permission preferences (see Figure 8). Figure 8a shows that participants under the age of 25 selected *Always Allow* permissions more frequently than those over 55 (35.0% vs. 25.5%), suggesting a generational difference in privacy attitudes. Additionally, participants with higher levels of education granted fewer data-sharing permissions overall (Figure 8b).

The self-reported metrics include AI tool familiarity, usage frequency, trust in AI tools, and privacy consciousness. We observe a significant shift in permission preference

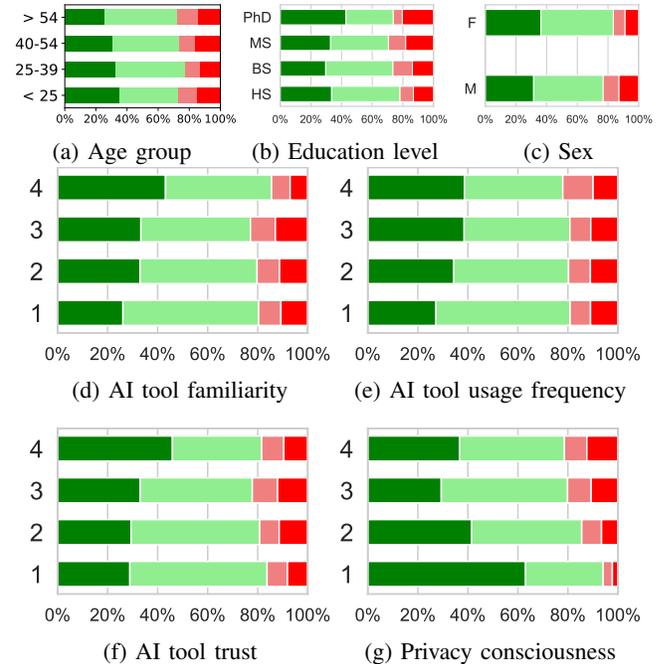


Figure 8: Group users according to their demographic information, which includes age group, education level, and sex, as well as their self-reported metrics such as AI tool familiarity, AI tool usage frequency, AI tool trust, and privacy consciousness level. For age groups, participants are divided into *below 25*, *25-39*, *40-55*, and *over 55*. Regarding education level, participants are categorized into four groups: high school or less (*HS*), bachelor’s (*BS*), master’s (*MS*), and PhD or other doctorate (*PhD*). For sex, *F* denotes female and *M* denotes male. Self-reported metrics are measured on a four-level scale (1-4) where a higher number indicates a greater level; for example, a privacy consciousness level of 4 signifies the highest concern for privacy.

distributions across different levels of these metrics. As shown in Figure 8d, participants more familiar with AI tools were more likely to select *Always Allow* and less likely to choose *Never Share*. Although AI usage frequency and trust (Figures 8e and 8f) do not show strong effects on overall permission preferences, those who used AI tools more often and reported higher trust levels tended to favor *Always Allow*. This suggests that greater engagement with AI may correlate with a preference for smoother user experiences, enabled by more permissive data-sharing settings.

For the privacy consciousness (Figure 8g), participants with higher levels of privacy consciousness were more likely to choose *Never Share*, reflecting a stronger desire to restrict data sharing. Interestingly, participants with medium levels of privacy consciousness selected *Always Allow* less often than those with high levels, but overall, they exhibited a more permissive data-sharing pattern compared to highly privacy-conscious individuals. While these self-reported measures may lack standardization or objectiv-

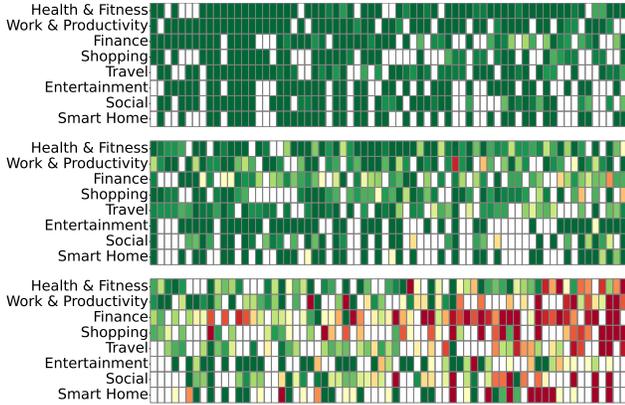


Figure 9: Heatmap of permission allowance rates across various domains for individual users. The figure is segmented into three distinct user groups. In each segmentation (user group), each column corresponds to a user and each row to a distinct domain (*Health & Fitness*, *Work & Productivity*, *Finance*, *Shopping*, *Travel*, *Entertainment*, *Social*, *Smart Home*). Darker green indicates a higher permission allowance rate, while darker red indicates a lower rate.

ity [14], and some participants may have inaccurate perceptions of their privacy attitudes, the overall trends provide useful insights into potential user data-sharing behaviors. In summary, our analysis indicates that demographic factors, AI-related attitudes, and privacy consciousness are associated with participants’ data-sharing preferences. These findings help us identify patterns that can be used to anticipate general user attitudes toward data-sharing permissions.

Appendix B. Analysis of User Permission Preferences Across Domains

Figure 9 provides insights into data-sharing permission preferences across domains among participants. It reveals a high degree of consistency in overall permission behaviors, as many participants exhibit uniform preferences across different domains, evident from extensive areas of homogeneous coloring. At the same time, variations in color intensity highlight specific domains, such as *Finance* and *Shopping*, where permission preferences significantly differ from more permissive areas like *Health & Fitness* and *Social* for some participants. These patterns emphasize that while participants generally maintain consistent permission preferences across domains, the distinct contexts of each domain can lead to observable variations in the data-sharing preferences of some participants.

Appendix C. CF Model Prediction Threshold

Since the CF model’s performance is highly sensitive to the prediction score threshold, we establish a baseline

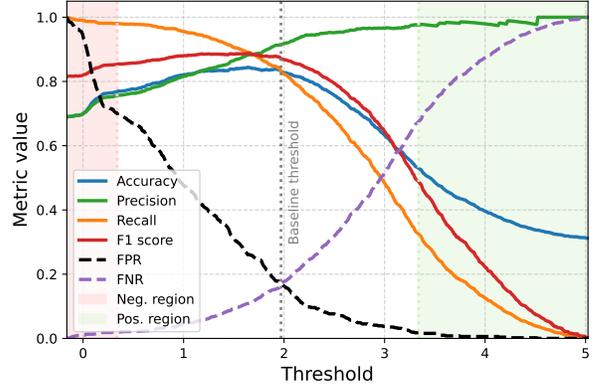


Figure 10: Collaborative filtering model metrics under different prediction score thresholds.

threshold for each fold in five-fold cross-validation by selecting the point where the FPR equals the FNR. Figure 10 illustrates the CF model’s metrics under varying thresholds for one of our test runs and shows the selected baseline threshold for reporting the model comparison results. For integrating the IC with CF, we use only CF predictions with high confidence. In the figure, the *negative region* and *positive region* define ranges of prediction scores that serve as additional contextual examples for the hybrid model.

Appendix D. Meta-Review

The following meta-review was prepared by the program committee for the 2026 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

D.1. Summary

This paper presents an automated permission management system for AI agents. The authors conducted a vignette-based survey of 205 users across a range of domains and scenarios to understand how users make permission decisions regarding data access and sharing in agentic systems. Building on these findings, the paper presents a preliminary design of a permission prediction model.

D.2. Scientific Contributions

- Independent Confirmation of Important Results with Limited Prior Research
- Establishes a New Research Direction

D.3. Reasons for Acceptance

- 1) The paper builds on a long line of work that seeks to understand factors influencing users' decisions in terms of sharing data and granting permissions to automated systems. Prior work in this space has focused on non-LLM-based personal assistants, and this paper obtains results for users' preferences in LLM-based agentic systems.
- 2) This paper is one of the first to investigate the problem of automating permission management for AI agents. It argues that, since automation is a key value proposition of AI agents, there is a need for permission management systems that can automatically make decisions on user's behalf. The paper establishes this new subfield of automated permission systems for AI agents, and presents one point in the design space of this problem.