PILLAR: LINDDUN Privacy Threat Modeling using LLMs

Majid Mollaeefar*, Andrea Bissoli*, Dimitri Van Landuyt†, and Silvio Ranise*‡
*Center for Cybersecurity, Fondazione Bruno Kessler (FBK), Trento, Italy
{mmollaeefar, abissoli, ranise}@fbk.eu
†KU Leuven, Leuven, Belgium
dimitri.vanlanduyt@kuleuven.be
‡Department of Mathematics, University of Trento, Trento, Italy

***Abstract*—The rapid evolution of Large Language Models (LLMs) has unlocked new possibilities for applying artificial intelligence across a wide range of fields, including privacy engineering. As modern applications increasingly handle sensitive user data, safeguarding privacy has become more critical than ever. To ensure robust data protection, potential threats must be identified and addressed early in the development process. Privacy threat modeling frameworks like LINDDUN offer structured approaches for uncovering these risks, yet they often require significant manual effort, expert knowledge, and detailed system information—making the process time-intensive and reliant on thorough analysis.**

To address these challenges, we introduce PILLAR (Privacy risk Identification with LINDDUN and LLM Analysis Report), a new tool that implements and automates the LINDDUN framework through LLM integration to streamline and enhance privacy threat modeling. PILLAR automates key parts of the LINDDUN process, such as generating DFDs from unstructured textual inputs (e.g. system descriptions), eliciting privacy threats, and risk-based threat prioritization. By leveraging the capabilities of LLMs, PILLAR can take natural language descriptions of systems and transform them into comprehensive threat models with limited input from users. Furthermore, PILLAR is capable of simulating multi-agent collaboration, allowing different LLM instances to play different contributor roles in a virtual threat modeling workshop.

Rather than merely reducing the workload on analysts, PILLAR shifts their involvement from repetitive, tedious tasks to more meaningful and impactful interventions—such as refining the scope of analysis or completing critical components like the DFD. This allows experts to focus on the aspects that truly matter for a robust threat modeling process while enhancing both efficiency and accuracy.

***Index Terms*—Privacy Threat Modeling, LINDDUN, Large Language Models**

1. Introduction

In today's digital landscape, privacy has become a paramount concern as applications increasingly handle sensitive user data. Ensuring robust privacy protection requires identifying and mitigating potential privacy threats during the early stages of system development. Privacy threat modeling frameworks, such as LINDDUN [25], provide structured approaches to uncover and address these threats. Thanks to their rigor structure, these frameworks often require significant manual effort, familiarity with the methodology itself (e.g., constructing a well-defined DFD or applying LINDDUN per interaction [22]), and domain-specific privacy knowledge (e.g., understanding concepts like non-repudiation, linkability, and attribution). This combination of requirements can be time-consuming and prone to oversight.

LINDDUN, one of the most mature privacy threat modeling approaches, involves creating and analyzing detailed Data Flow Diagrams (DFDs) and other system descriptions to identify potential privacy risks. While effective, this process is often cumbersome and depends heavily on the accuracy and completeness of the input provided by users. Additionally, these methods can struggle with the prioritization of identified threats, leading to an overwhelming list of potential issues without clear guidance on which to address first.

To address these challenges, we present **PILLAR** (Privacy risk Identification with the LINDDUN and LLM Analysis Report), a novel tool that integrates Large Language Models (LLMs) with the LINDDUN framework to automate and enhance the privacy threat modeling process. PILLAR is designed to simplify the identification and analysis of privacy threats by automating key aspects of the LINDDUN methodology, such as the generation of DFDs, the categorization of threats, and the prioritization of risks. By leveraging LLMs, PILLAR can process natural language descriptions of systems to produce detailed threat models with minimal user input, reducing the burden on developers and researchers.

A number of LLM integration efforts have been performed mainly in the area of security threat modeling. These efforts, however, fail to mimic one of the crucial tenets of threat modeling [1], [17] i.e. that multi-stakeholder and multi-perspective collaboration (typically in workshop sessions) is required to address different perspectives. To address this discrepancy, PILLAR implements an innovative approach to threat modeling by simulating multi-agent collaboration among virtual experts, each focused on different aspects of privacy threats. In PILLAR, virtual agents, such as a privacy expert or a developer, communicate and debate privacy risks using distinct but complementary perspectives. This structure reflects the *cooperative communication* paradigm outlined by Guo et al. (2024) [13], where multiple agents cooperate to achieve a shared goal through interactions and knowledge sharing. By simulating multiple rounds of communi-

cation between virtual agents, PILLAR tries to reduce the likelihood of overlooking critical privacy risks, mirroring the real-world collaborative efforts between stakeholders in privacy threat modeling. This approach enables the tool to capture a broader range of potential threats by allowing agents to deliberate and adjust their insights iteratively, leading to more thorough risk assessments.

In this paper, we present the design and implementation of PILLAR, which leverages LLM technology to introduce several key innovations that enhance and streamline the privacy threat modeling process through automation:

Automated Application Description. Users can provide a natural language description of their application, specifying details such as application and data type, policies, etc. where the automated intake of such information reduces manual effort.

DFD Management. Enables users to generate DFD from various input types (e.g., textual or visual) while providing capabilities such as editing and exporting DFDs in an interoperable format (`CSV`) for later use.

LINDDUN Analysis. It automates three variants of LINDDUN, acronym-based brainstorm (SIMPLE), LINDDUN GO and LINDDUN PRO.

Multi-agent Collaboration. It simulates a multi-agent collaboration among virtual experts each focused on different aspects of privacy threats.

Impact Assessment and Controls. The tool generates an automated assessment of identified threats and recommends privacy controls based on established privacy patterns.

Comprehensive Reporting. It generates detailed reports summarizing threats, impacts, and control measures, aiding in documentation and compliance.

These contributions demonstrate PILLAR's potential to automate, simplify, and enhance the privacy threat modeling process, making it more accessible and efficient for developers, privacy experts, and security researchers alike. We validate PILLAR by comparing its outputs to a manual evaluation of privacy threats in the context of a contact tracing management application.

A critical research question driving this study is: *What value does PILLAR bring to privacy threat modeling? Specifically, does it enhance efficiency, scalability, or repeatability compared to traditional manual methods?* Addressing this question allows us to evaluate PILLAR's contributions to privacy threat modeling compared to traditional, manual methods. The remainder of this paper is organized as follows: Section 2 provides the necessary background on threat modeling and Large Language Models. Section 3 details the architecture and functionality of the PILLAR tool. Section 4 demonstrates the effectiveness of PILLAR by applying it to a case study, and presenting evaluation results together with a discussion. Related work is discussed in Section 5. Finally, Section 6 concludes with a discussion of future work and potential enhancements.

2. Background

In this section, we focus on two primary areas: the importance of threat modeling in cybersecurity and privacy, and the application of Large Language Models to enhance the threat modeling process. By combining these methodologies, we aim to address the challenges in identifying and mitigating privacy risks efficiently and effectively.

2.1. Threat Modeling

Threat modeling is a foundational practice in privacy management, providing a structured methodology to identify, assess, and mitigate potential risks to personal data within a system and to support compliance with regulations such as the General Data Protection Regulation (GDPR). It is crucial for several reasons: First, it enables organizations to systematically identify areas where personal data might be exposed or misused by examining how information is collected, processed, stored, and shared. Second, threat modeling helps prioritize risks based on both their potential impact on data subjects and the likelihood of occurrence, ensuring that resources are allocated to address the most pressing privacy concerns. Third, it fosters clear communication among various stakeholders—including developers, privacy teams, and management—ensuring that all parties understand the risks and the corresponding mitigation strategies. Moreover, threat modeling supports compliance with regulatory requirements, as many legal frameworks mandate comprehensive assessments of potential data protection issues. Finally, threat modeling is not a one-time effort; it is an ongoing process that should be integrated into the software development life cycle (SDLC), allowing for continuous evaluation and adaptation to emerging privacy challenges.

There are several established methodologies for conducting threat modeling [23], [24], each offering a distinct approach. Among these various methods, two of the most well-known system-centric approaches are STRIDE [1] for security and LINDDUN [25] for privacy.

2.2. The LINDDUN Framework

The LINDDUN framework is a comprehensive tool for conducting privacy threat modeling, allowing organizations to systematically identify and mitigate privacy risks throughout the software development lifecycle. Each letter of "LINDDUN" stands for a privacy threat type (Linking, Identifying, Non-Repudiation, Detecting, Data Disclosure, Unawareness and Unintervenability, and Non-Compliance). LINDDUN has received recognition as one of the most extensive privacy threat modeling frameworks and is well known and used both in academia and industry [18], [19] and is considered a recommended practice [20]. LINDDUN operates in two main phases:

Problem Space. This phase involves creating DFDs, mapping privacy threats to elements within the DFD, and identifying potential threat scenarios.

Solution Space. In this phase, identified threats are prioritized, mitigation strategies are elicited, and corresponding privacy-enhancing technologies (PETs) are selected.

To serve different needs, LINDDUN comes in various flavors. These approaches vary in complexity and comprehensiveness, ranging from lean to in-depth analysis.

LINDDUN framework can be applied through three different methods[1], GO, PRO, and MAESTRO[2].

LINDDUN GO is a toolkit in the sense that it offers both privacy threat information to use and guidelines on how to apply it in a systematic way. It consists of a set of threat types cards that describe the most common privacy threats for each threat category [14]. This light version of the framework reduced the number of threats to 33 by combining related threat types and discarding less important ones. It is presented as a more accessible, lightweight version with more consideration on its usability aspects including effort and ease of use.

LINDDUN GO introduces an extended and structured version of LINDDUN's threat trees[3], documenting them as threat type cards. Each card follows a consistent template. This template includes essential elements such as a title, hotspots (locations within the system where threats occur), threat source (origin of the threat), summary, elicitation questions (to assess threat applicability), examples, consequences/impact, additional information, a unique identifier, and a corresponding LINDDUN category.

Executing LINDDUN GO is a straightforward and collaborative process[4]. Each participant takes turns drawing a random threat type card from a pile and attempts to elicit corresponding threats. Other participants then fill in any missing threats related to the card. If no threats are found, the next participant draws a card, and the process repeats. To elicit threats, participants read the drawn threat type card and assess whether the threat is applicable to each system component corresponding with the hotspot described on the card. They can answer the two applicability questions on the card to aid this assessment.

LINDDUN PRO also known as "LINDDUN-per-interaction" [22] is a systematic and comprehensive approach for modeling privacy threats, aimed at identifying a wide spectrum of privacy threats. Unlike LINDDUN GO, which prioritizes ease of use and collaboration, LINDDUN PRO emphasizes a thorough and structured methodology, making it suitable for more advanced threat modeling scenarios where precision and exhaustiveness are paramount.

The starting point of a LINDDUN PRO analysis is a Data Flow Diagram (DFD), which serves as the system abstraction for identifying potential threats. DFDs represent the data flows, stores, processes, and external entities within a system, acting as a visual guide for mapping privacy threats to specific system elements. By systematically analyzing each interaction in the DFD, LINDDUN PRO ensures a detailed examination of how data is processed, stored, and shared, enabling a robust identification of privacy risks.

One of the key strengths of LINDDUN PRO is its extensive knowledge support, which includes detailed privacy threat types, comprehensive privacy threat trees with examples, evaluation criteria, and a mapping table to connect DFD elements with corresponding privacy threat characteristics. These resources allow practitioners to methodically explore potential risks by applying threat trees to specific data flows, processes, or data stores. Threat trees provide a structured breakdown of each threat category, helping users to identify root causes, contributing factors, and potential impacts of privacy risks.

The method encourages systematic iteration over all DFD elements and their interactions, ensuring no potential privacy risk is overlooked. However, the level of detail and the exhaustive nature of the process make LINDDUN PRO resource-intensive, requiring significant time and expertise. Therefore, many practitioners only brainstorm over the LINDDUN acronyms as a first exploration.

2.3. Privacy Patterns

In addition to structured frameworks like LINDDUN, privacy patterns[5] have emerged as a useful tool for enhancing privacy practices during system design. Privacy patterns are reusable solutions to common privacy-related challenges, such as data minimization, user consent, and transparency. By adopting these patterns, developers can address specific privacy concerns and improve compliance with privacy regulations. Moreover, the use of privacy patterns encourages a proactive approach to privacy, embedding best practices into system designs from the outset.

2.4. Large Language Models (LLMs)

LLMs represent a significant advancement in natural language processing and have shown potential for a variety of applications in the field of security. LLMs are built on transformer architectures that allow them to process vast amounts of text data and learn complex linguistic patterns. These models are typically trained on massive datasets, comprising billions of words, enabling them to generate coherent and contextually relevant text in response to textual user prompts.

LLMs are particularly well-suited for tasks involving the analysis of natural language system documentation, the generation of code or system models, and assisting in various analytical and automation tasks. Their ability to understand and generate human-like text makes them a valuable tool for automating processes that traditionally require expert input, such as privacy threat modeling.

However, the use of LLMs is not without challenges. These models can inherit biases from their training data, which raises concerns about fairness and accuracy in their outputs. Additionally, while LLMs excel at generating coherent responses, they may not always produce technically accurate or unbiased results without careful prompt engineering and refinement [21].

3. Tool Architecture

The PILLAR tool provides (LLM) support in four essential phases of the threat modeling process: *System Description*, *Threat Elicitation*, *Impact Assessment* and *Controls*, and *Report Creation*. Each phase is described below. For a detailed overview, refer to the tool architecture documentation provided in a complementary material

1. https://linddun.org
2. Currently unavailable.
3. https://linddun.org/threat-trees
4. https://linddun.org/go-getting-started/
5. https://privacypatterns.org

document[6]. PILLAR[7] is implemented as a web application offering a user-friendly interface for privacy threat modeling. Python was chosen for its prototyping speed and widespread use in machine learning development. It integrates LLMs such as OpenAI[8], Gemini[9] and Mistral[10]. OpenAI serves as the primary provider, offering robust developer tools and cost-effective models like *gpt-4o-mini* for improved response quality. It also supports OpenAI's structured JSON output for reliable formatting.

3.1. System Description

The user needs to describe the system to be analyzed, such that the subsequent threat elicitation can be performed. Naturally, the more detailed information provided about the system, including specific technical details and assumptions about the data, the more likely the resulting output will be of higher quality. Alternatively, users can define the system through a DFD, which supports manual editing, graph visualization, and generation via LLMs using textual descriptions or uploaded DFD images[11]. Both allow content to be downloaded or uploaded as `csv` files for external editing or future use. These input methods minimize user effort while enabling precision and refinement.

3.2. Threat Elicitation

Once the system description is provided, threat elicitation follows the LINDDUN framework. PILLAR supports three approaches: (i) SIMPLE, which is based on acronyms, and (ii) LINDDUN GO—both of which can operate using either a textual system description, a DFD, or a combination of both. In contrast, (iii) LINDDUN PRO requires the DFD as an input. Below, these three methods are described.

3.2.1. SIMPLE. It is a basic zero-shot threat elicitation where the LLM uses the system description provided and identifies threats with a focus on each of LINDDUN's threat categories. The output can provide initial insight on the types of threats to be aware of while minimizing the required effort from the user.

3.2.2. LINDDUN GO. For this method, two modes are supported: single-agent and multi-agent. In the single-agent simulation, each card's description and information is provided to the LLM, together with the system description. The LLM's task is to determine whether or not the threat contained in the card is applicable to the system and the reason for either decision. The output specifies which threats are relevant and why they should be taken into consideration. To effectively simulate this method, we adopted a structured, iterative multi-agent process that mimics the gamified collaborative brainstorming nature of the approach. For each card, different LLM agents are spawned, each with a different prompt that suggests its area of expertise and focus, just like real-life members of a privacy threat modeling team. Each of them carries out the single-agent analysis focusing on the aspects important to their own area. During the multi-agent analysis, if specified, LLM agents are randomly selected from the different providers (i.e., OpenAI, Mistral, and Google Gemini). The process of simulating LINDDUN GO with multi-agent integration is shown in Fig. 1, where it consists of the following steps:

Card Formalization and Agent Assignment *(shown in the top-left side of Fig. 1)*: The entire set of LINDDUN GO cards has been consolidated into a single JSON file, which includes the threat name, threat type, and the corresponding questions to be asked. Additionally, for each card, we have predefined which agent roles will be most competent to assess the associated threats. For example, for the "Profiling users" (L05) threat card (part of the Linking threat category), two specific questions must be posed:

- *Are there patterns derivable from the data?*
- *Can (new) personal data be inferred from the linked data points?*

To provide well-informed responses, we designate the *Cybersecurity Expert, Software Developer, and Data Protection Officer (DPO)* as the most competent agents[12] to evaluate these concerns. This assignment serves two purposes; i) it ensures that only qualified agents provide reasoning, avoiding non-relevant discussions, and ii) by reducing unnecessary debates, the overall computational cost (e.g., token usage) is minimized.

Iterative Threat Elicitation Process *(shown in the top-right side of Fig. 1)*: At the start of each round, a card is selected, and its content, along with the application description provided by the user, is supplied as input to structure the LLM prompts for each agent. Each agent analyzes the threat and produces a reasoned response, evaluating the presence of the threat based on their expertise. At the end of each round, the agents' responses are gathered into a summary report, which is shared with all agents as additional input for the next round. This iterative sharing of responses enables agents to refine their reasoning, similar to how real-life experts exchange insights during brainstorming sessions.

Final Judgment Phase: Once the iterative rounds have completed, the collected outputs are passed to a *Judge Agent* (shown at the bottom of Fig. 1) which corresponds to the role of the threat modeling workshop moderator. This agent evaluates the collective reasoning from all agents and delivers a final verdict on the existence and relevance of the threat.

The PILLAR tool is designed so that the user can specify the number of cards to extract from the LINDDUN GO's deck and whether or not to carry out a multi-agent simulation.

3.2.3. LINDDUN PRO. This method requires a DFD of the system under analysis. Next, the user selects a specific

6. https://tinyurl.com/39atws6m
7. https://github.com/stfbk/PILLAR
8. https://openai.com
9. https://gemini.google.com
10. https://mistral.ai
11. As this paper focuses on the threat elicitation, prioritization and mitigation, we do not further detail how a DFD is constructed and extracted from images. The reader is strongly encouraged to experiment with the tool to understand these capabilities.
12. For brevity, the definitions of agents and their tasks are not included here; readers can refer to the code repository in the file `./llms/prompts.py` for further details.

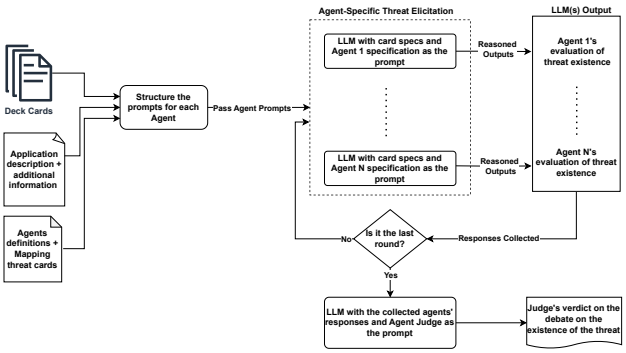

Figure 1. Process view of PILLAR's LINDDUN GO implementation with multi-agent integration.

interaction in the DFD to analyze and chooses one or more LINDDUN threat types for assessment. A description of the specific data flow also needs to be added for the LLM to understand how data is handled throughout the system. Once this information is provided, the user can generate a probable threat of the specified threat type which is linked to a specific location within the interaction (source, data flow, or destination). The process can be repeated many times, with different threat types and other DFD interactions.

Under the hood, PILLAR uses the LINDDUN PRO mapping table[13] to decide whether a threat assessment is applicable for a certain DFD edge. Furthermore, the LLM receives a textual representation version of LINDDUN's threat trees including the examples, etc., and bases its analysis on the privacy threat characteristics (leaf nodes of the trees). When presenting the threat, the output also refers to the specific threat tree node upon which the identified threat is based.

3.3. Impact Assessment and Controls

After threats have been elicited using one of PILLAR's methodologies, the Impact Assessment phase enables a structured evaluation of their potential privacy implications. Users can import the identified threats and systematically assess their relevance for inclusion in the final analysis. Leveraging LLM capabilities, PILLAR facilitates the generation of preliminary impact assessments for each threat. These assessments serve as an initial analytical foundation and can be refined further by users to align with specific system contexts and risk tolerances.

Control measures for the identified threats are generated using *privacy patterns*. Initially, the LLM receives the threat, system description, and a structured summary of privacy patterns extracted from their official sources and integrated into PILLAR's knowledge base. Based on this input, the LLM generates a list of potentially relevant patterns for addressing the threat. Then, a more detailed request is sent to the LLM which includes a complete description of the selected privacy patterns. Based on these detailed inputs, the LLM then performs a further selection among them, as well as offering reasons for each pattern's relevance and guidelines on how to implement it within the system.

13. https://tinyurl.com/p8chbncc

4. Validation and Demonstration of PILLAR

We validate the effectiveness of PILLAR by applying it to a real-world use-case scenario. During this project, the main objective was to identify and assess the potential security and privacy threats that could arise in the system, using manual threat modeling techniques.

4.1. Use-Case Scenario Description

Trace4 (anonymised name) is an enterprise contact tracing solution created in the context of a European project to contribute to COVID-19 pandemic management. It is designed to address the challenges posed by this pandemic in enterprise settings. It aims to monitor and enforce physical distancing, analyze risks associated with potential outbreaks, and manage such outbreaks effectively. The solution operates within a scenario where the primary actors include the COVID-19 safety responsible person (service provider), end-users (employees), and the team leader with authorized access to contact tracing data. Communication channels, such as gateways and Bluetooth Low Energy channels, are used for data collection and transfer.

As part of the Trace4 project, we conducted a GDPR-mandated DPIA to identify security and privacy risks in data processing, storage, and sharing. To support this, we distributed a questionnaire to project partners, gathering insights into privacy concerns at different stages of the scenario. Complementary to the DPIA, we also performed a privacy threat modeling exercise, identifying threat scenarios within the Trace4 solution at the basis of the LINDDUN threat types.

The privacy threats identified are summarized in Table 1. These threats encompass risks such as user tracking, unauthorized data sharing, and lack of transparency, each posing significant privacy challenges to the system's compliance and trustworthiness. We consider this threat modeling outcome to be the "ground truth" baseline for comparison in this validation.

4.2. Application of PILLAR to Trace4

To validate PILLAR's capabilities in a practical setting, we applied it to the Trace4 use case. For this validation, we specified the LLM provider and model used for this analysis—specifically, *OpenAI GPT-4o mini* and *Mistral large-dataset*. For brevity, we provide a detailed analysis of these methods in the complementary material document. Readers are encouraged to refer to this document, which includes the PILLAR tool architecture, the detailed input provided to PILLAR for the Trace4 scenario, a comparison of results using the SIMPLE method, a multi-agent debate simulating the LINDDUN GO approach, and a report generated by PILLAR using the LINDDUN PRO method. Below, we discuss our observation on privacy threat modeling for the Trace4 scenario by using PILLAR.

4.3. Results

PILLAR provides three distinct approaches for privacy threat modeling: the SIMPLE method, the LINDDUN

TABLE 1. MANUALLY IDENTIFIED PRIVACY THREAT SCENARIOS.

T	Threat Scenarios	Consequences
I, D	MT1- An adversary equipped with a Bluetooth Beacon Tracker can observe tokens, and in the case of token IDs not changing over time, the attacker can re-identify token holders.	Tracking users, identifying users, profiling users' behavior, learning about places.
I, D	MT2- An attacker can eavesdrop on the network traffic by setting up their device close to the gateways when data is uploaded on gateways.	Identifying users.
NC	MT3- The data is stored for longer, which increases the chance of data abuse and decreases its security.	Violates storage limitation and data minimization.
I, L	MT4- Identifying an entity from a set of collected data, e.g., in our case, identifying positive cases.	Re-identify users.
UU	MT5- Users' information is shared with a third party or submitted to the health authority without their explicit consent.	Violates lawfulness, fairness, and transparency.
UU	MT6- Lack of sufficient and complete description of the service and operation details (such as data flows, data storage location, transmission methods, etc.) and their impacts on users' data.	Lack of transparency and non-compliance with the law.
UU	MT7- Users cannot submit correction requests (in the case of wrongly recorded contacts) that need to be evaluated by the system administrator. There is no implemented procedure in the system to allow the users to notify the system administrator to rectify, erase, or block the wrongly registered contacts. For instance, in undefined events, if wrong contacts are uploaded (registered) in the system, it causes the contact tracing network to be created incorrectly and results in incorrect notifications.	Lack of control and inability to rectify or erase wrongly registered contacts. Loss of trust in the system, impact on the contact tracing network.
DD	MT8- Unauthorized access to the local stores.	Information disclosure.

GO, and LINDDUN PRO. This section discusses and compares the results obtained using the first two methods.

Performance of the SIMPLE Method. Fig. 2 shows a privacy threat modeling breakdown using this method for the Trace4 scenario. The SIMPLE method identified a total of **21** threat scenarios. Among these, **13** threats were relevant and aligned with those manually identified in Table 1, demonstrating strong consistency with human-led threat modeling. This resulted in a Recall of **100%**, indicating that all threats found in the manual analysis were also detected by PILLAR.

Additionally, PILLAR uncovered **4** new relevant threats that were not explicitly considered in the manual analysis. These threats belonged to "L, NR, and UU" categories. Some of these introduce accountability risks, while others highlight potential transparency gaps in privacy processes. This expansion demonstrates the ability of the tool to go beyond manual assessments by identifying additional threats that we did not discover in our threat elicitation. Including these new insights, PILLAR achieves an overall Precision of **85.71%**.

However, **4** threats generated by PILLAR were deemed irrelevant to the system context, leading to a False Positive Rate of **19.05%**. For instance, one such threat scenario states: *Employees are not trained on data privacy policies, leading to unintentional breaches of data handling procedures.* While this could be a valid concern

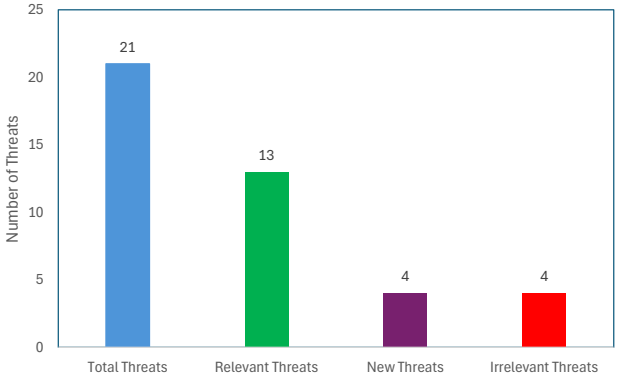

Figure 2. PILLAR threat modeling breakdown using the SIMPLE method

in some contexts, it is not applicable to the Trace4 solution, where users merely carry wearable tokens and do not actively handle or process sensitive data. This reflects a common trade-off in automated threat modeling, where increased sensitivity to potential threats can sometimes introduce irrelevant scenarios. These findings directly address our *research question* mentioned in Section 1. PILLAR clearly demonstrates its ability to expand the scope of privacy threat discovery by identifying risks that were overlooked in manual analysis.

Insights from the LINDDUN GO Multi-Agent. To explore this method, we simulated a multi-agent discussion using PILLAR, focusing on the privacy threat of *Non-Repudiation of Storage*, one of the LINDDUN GO deck cards. For a detailed review of the debate among agents, we refer the reader to the complementary material. However, a short summary of this analysis is as follows:

The debate unfolded over three rounds, where multiple agents (i.e., Domain Expert, Legal Expert, and CISO) provided their assessments. Initially, opinions were divided—some agents concluded that the threat was not present due to the lack of explicit mention of digital signatures, while others argued that the absence of such mechanisms could pose a risk. As the debate progressed, the agents reconsidered their positions by incorporating insights from one another. By the final round, all agents reached a consensus that the threat was indeed present due to the lack of digital signatures or non-repudiation mechanisms.

4.4. Discussion

We have validated PILLAR in two complementary experiments. In the first, we have shown the capability and effectiveness of approximating expert threat evaluation outcomes using the SIMPLE open-ended approach. In the second, we report on the multi-agent implementation of LINDDUN GO, and we illustrate the value of simulating the gamified method between different LLM agents, allowing them to provide arguments and discuss.

While these results already evidence the value of this approach, more extensive empirical evaluation of the different PILLAR capabilities is considered future work.

5. Related Work

The integration of LLMs into the field of cybersecurity has garnered considerable attention due to their ability to process and analyze vast datasets [21]. As cyber threats become more sophisticated, the cybersecurity domain increasingly turns to these advanced models to strengthen defenses. Cybersecurity professionals continually seek innovative solutions to implement robust policies and enhance technological safeguards [2]. These efforts are essential for preventing the unauthorized disclosure of sensitive information, unauthorized access, and various forms of data manipulation [3]. The ability of LLMs to adapt and scale is particularly valuable for addressing the growing complexity of cyber threats. Some applications of LLMs in cybersecurity include network intrusion detection, Cyber Threat Intelligence (CTI), threat modeling, vulnerability detection, malware analysis, and phishing prevention.

In the context of vulnerability detection, Liu et al. [9] proposed LATTE, which combines LLMs with automated binary taint analysis. LATTE overcomes the limitations of traditional taint analysis, which often requires manual customization of taint propagation and vulnerability inspection rules. Phishing and scam detection is another area where LLMs demonstrate significant utility. Labonne et al. [8] highlighted the effectiveness of LLMs in spam email detection, showcasing their superiority over traditional machine learning approaches. Cambiaso et al. [7] presented an innovative study suggesting that LLMs can mimic human interactions with scammers in an automated yet meaningless manner, wasting scammers' time and resources, and thereby reducing the impact of scam emails.

One significant application of LLMs in network security is web fuzzing. Liang et al. [4] proposed GPTFuzzer, which uses an encoder-decoder architecture to generate effective payloads for web application firewalls. It specifically targets vulnerabilities such as SQL injection, cross-site scripting, and remote code execution by generating fuzz test cases. To detect network traffic anomalies, Liu et al. [10] developed a method to detect malicious URLs by leveraging LLMs to extract hierarchical features. This work extends the use of LLMs in intrusion detection tasks to the user level, demonstrating their generality and effectiveness in intrusion and anomaly detection.

As cyber threats grow in complexity, traditional CTI methods struggle to keep pace. AI-based solutions, including LLMs, offer an opportunity to automate and enhance several tasks, ranging from data ingestion to resilience verification. LLMs have shown promise in generating CTI from various sources, including network security texts (e.g. books, blogs, news) [6], generating structured reports from unstructured data [12], and extracting intelligence from network security entity graphs [11]. Aghaei et al. [5] introduced CVEDrill, which generates priority recommendation reports for cybersecurity threats and predicts their potential impact.

As for threat modeling, Gadyatskaya and Papuc [15] utilized GPT-3.5 to automatically create attack trees, a universal method of threat modeling that represents threats from an attacker's perspective. Their findings showed that GPT-3.5 includes enough information about cyber threats and their relationships to construct attack trees, which improves the automation level of threat identification and the discovery of unknown threats. However, limitations were identified, including the inability to identify detailed operators. Yang et al. [16] introduced ThreatModeling-LLM, a novel and adaptable framework that automates threat modeling for banking systems using LLMs. The framework operates in three stages: i) dataset creation, ii) prompt engineering, and iii) model fine-tuning. They first generated a benchmark dataset using the Microsoft Threat Modeling Tool. Then, they applied a chain of thought and optimization by prompting on the pre-trained LLMs to optimize the initial prompt. As the last stage, fine-tune the LLM using Low-Rank Adaptation based on the benchmark dataset and the optimized prompt to improve the threat identification and mitigation generation capabilities of pre-trained LLMs. STRIDE-GPT[14] is another advanced tool that generates threat models and attack trees for a given application based on the STRIDE methodology. AttackGen[15] is a cybersecurity incident response testing tool that leverages the power of large language models and the comprehensive MITRE ATT&CK framework[16]. The tool generates tailored incident response scenarios based on user-selected threat actor groups and your organization's details. In complement and contrast to these emerging LLM-based threat modeling approaches, PILLAR mimics or simulates the workshop-oriented nature of threat elicitation efforts in the multi-agent implementation of LINDDUN GO. To the best of our knowledge, this is the first effort to automate this specific aspect of the threat modeling process.

6. Conclusion

In this paper, we introduced **PILLAR**, an LLM-powered privacy threat modeling tool that integrates the LINDDUN framework to automate and enhance privacy risk identification. By automating processes such as threat identification, and risk prioritization, PILLAR eases the burden on developers and privacy experts, particularly those operating with limited resources or specialized expertise. This automation aims to improve both the *efficiency* and *accuracy* of threat modeling.

Beyond these automated capabilities, PILLAR incorporates *multi-agent collaboration*, allowing large language models to simulate expert deliberation. This approach refines privacy threat analyses by generating more robust outcomes through simulated discussions and consensus-building. However, as with any generative AI solution, expert oversight remains essential to validate the tool's outputs and mitigate the risk of incomplete or inaccurate threat assessments.

While PILLAR represents significant progress in privacy threat modeling, several challenges and future directions remain:

Balancing Assistance and Expertise. Rather than replacing human analysts, LLMs function best as *assistants*, handling tasks such as generating structured reports and refining threat descriptions. This helps

14. https://github.com/mrwadams/stride-gpt
15. https://github.com/mrwadams/attackgen
16. https://attack.mitre.org/

privacy experts concentrate on high-level decision-making and complex risk judgments.

Enabling Continuous, Interactive Threat Modeling. PILLAR's automation facilitates frequent and iterative updates to privacy threat models as system designs evolve. Future work could explore interactive querying—e.g., *"Is the data structure sufficiently de-identified?"*—and instantaneous, context-rich feedback.

Rethinking the Role of DFDs. While DFDs have traditionally helped human analysts visualize data flows, LLMs can often infer system behavior from textual descriptions alone. Future research might determine whether bypassing DFD generation can streamline processes without sacrificing accuracy.

Enhancing LLM Output Reliability. As a generative AI application, PILLAR must address the risk of inaccurate or incomplete results. Integrating *Retrieval-Augmented Generation (RAG)* can anchor the model's outputs to up-to-date, authoritative information sources, thereby improving confidence in privacy threat assessments.

Expanding Interoperability. To further strengthen privacy assessments, PILLAR can integrate with other security and privacy tools. Seamless data exchange would allow organizations to benefit from a broader ecosystem of risk assessment frameworks. Furthermore, the ability to interoperate with other development tools is a prerequisite for integration in CI/CD [26], which is in turn essential for continuous threat awareness in the development lifecycle.

Agent Deliberation. Additionally, our multi-agent approach can be enhanced by encouraging structured debates among agents. Recent research[17] suggests that agent deliberation can improve the precision of generative models through simulated argumentation and consensus-building. Incorporating these mechanisms into PILLAR's multi-agent framework could foster more rigorous threat validation and improve the robustness of the final results.

Looking ahead, our primary objectives include further evaluating and refining the accuracy of threat generation, expanding PILLAR's interoperability with other tools, and exploring advanced LLM techniques—such as expanded RAG implementations—to further reduce manual input. By continuously improving these aspects, PILLAR can evolve into a comprehensive tool that helps organizations enhance their data protection practices and meet regulatory obligations, including compliance with regulations such as the GDPR.

Acknowledgment

This work was partially supported by the project SERICS (PE00000014), under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

[1] A. Shostack, Threat Modeling: Designing for Security, John Wiley & Sons, 2014.

17. https://tinyurl.com/uzhc3mj8

[2] R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial intelligence for cybersecurity: Literature review and future research directions," Information Fusion, vol. 97, p. 101804, 2023, Elsevier.

[3] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu, "Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond," ACM Trans. Knowledge Discovery from Data, vol. 18, no. 6, pp. 1–32, 2024, ACM New York, NY.

[4] H. Liang, X. Li, D. Xiao, J. Liu, Y. Zhou, A. Wang, and J. Li, "Generative pre-trained transformer-based reinforcement learning for testing web application firewalls," IEEE Trans. Dependable Secure Comput., vol. 21, no. 1, pp. 309–324, 2023, IEEE.

[5] E. Aghaei, E. Al-Shaer, W. Shadid, and X. Niu, "Automated CVE analysis for threat prioritization and impact prediction," arXiv preprint arXiv:2309.03040, 2023.

[6] E. Aghaei, X. Niu, W. Shadid, and E. Al-Shaer, "Securebert: A domain-specific language model for cybersecurity," in Proc. Int. Conf. Security Privacy Commun. Systems, 2022, pp. 39–56, Springer.

[7] E. Cambiaso and L. Caviglione, "Scamming the scammers: Using ChatGPT to reply mails for wasting time and resources," arXiv preprint arXiv:2303.13521, 2023.

[8] M. Labonne and S. Moran, "Spam-t5: Benchmarking large language models for few-shot email spam detection," arXiv preprint arXiv:2304.01238, 2023.

[9] P. Liu, C. Sun, Y. Zheng, X. Feng, C. Qin, Y. Wang, Z. Li, and L. Sun, "Harnessing the power of LLM to support binary taint analysis," arXiv preprint arXiv:2310.08275, 2023.

[10] R. Liu, Y. Wang, H. Xu, Z. Qin, Y. Liu, and Z. Cao, "Malicious URL detection via pretrained language model guided multi-level feature attention network," arXiv preprint arXiv:2311.12372, 2023.

[11] F. Perrina, F. Marchiori, M. Conti, and N. V. Verde, "Agir: Automating cyber threat intelligence reporting with natural language generation," in Proc. IEEE Int. Conf. Big Data (BigData), 2023, pp. 3053–3062, IEEE.

[12] G. Siracusano, D. Sanvito, R. Gonzalez, M. Srinivasan, S. Kamatchi, W. Takahashi, M. Kawakita, T. Kakumaru, and R. Bifulco, "Time for action: Automated analysis of cyber threat intelligence in the wild," arXiv preprint arXiv:2307.10214, 2023.

[13] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," arXiv preprint arXiv:2402.01680, 2024.

[14] K. Wuyts, L. Sion, and W. Joosen, "LINDDUN GO: A lightweight approach to privacy threat modeling," in Proc. IEEE European Symp. on Security and Privacy Workshops (EuroS&PW), 2020, pp. 302–309, IEEE.

[15] O. Gadyatskaya and D. Papuc, "ChatGPT knows your attacks: Synthesizing attack trees using LLMs," in Data Science and Artificial Intelligence, C. Anutariya and M. M. Bonsangue, Eds., 2023, pp. 245–260, Springer Nature Singapore.

[16] S. Yang, T. Wu, S. Liu, D. Nguyen, S. Jang, and A. Abuadbba, "ThreatModeling-LLM: Automating threat modeling using large language models for banking system," arXiv preprint arXiv:2411.17058, 2024.

[17] Z. Braiterman, A. Shostack, J. Marcil, S. de Vries, I. Michlin, K. Wuyts, R. Hurlbut, B. S. E. Schoenfield, F. Scott, M. Coles, C. Romeo, A. Miller, I. Tarandach, A. Douglen, M. French, "Threat Modeling Manifesto," Nov. 2020.

[18] S. Wairimu, L. H. Iwaya, L. Fritsch, and S. Lindskog, "On the evaluation of privacy impact assessment and privacy risk assessment methodologies: A systematic literature review," IEEE Access, 2024, IEEE.

[19] E. D. Canedo, I. N. Bandeira, A. T. S. Calazans, P. H. T. Costa, E. C. R. Cançado, and R. Bonifácio, "Privacy requirements elicitation: A systematic literature review and perception analysis of IT practitioners," Requirements Eng., vol. 28, no. 2, pp. 177–194, 2023, Springer.

[20] International Organization for Standardization (ISO), “ISO/IEC TR 27550:2019: Information technology — Security techniques — Privacy engineering for system life cycle processes,” Sep. 2019.

[21] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, “A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly,” High-Confidence Computing, p. 100211, 2024, Elsevier.

[22] L. Sion, K. Wuyts, K. Yskout, D. Van Landuyt, and W. Joosen, “Interaction-based privacy threat elicitation,” in Proc. IEEE European Symp. on Security and Privacy Workshops (EuroS&PW), 2018, pp. 79–86, IEEE.

[23] W. Xiong and R. Lagerstr"om, “Threat modeling—A systematic literature review,” Computers & Security, vol. 84, pp. 53–69, 2019, Elsevier.

[24] K. Tuma, G. Calikli, and R. Scandariato, “Threat analysis of software systems: A systematic literature review,” Journal of Systems and Software, vol. 144, pp. 275–294, 2018, Elsevier.

[25] K. Wuyts, D. Van Landuyt, L. Sion, and W. Joosen, “LINDDUN privacy threat types,” Aug. 2023.

[26] D. Van Landuyt, L. Sion, W. Philips, and W. Joosen, “From automation to CI/CD: A comparative evaluation of threat modeling tools,” in Proc. IEEE Secure Development Conf. (SecDev), 2024, pp. 35–45, IEEE.