The Right Tool for the Job: On the Selection of Mitigations for GenAI Privacy Threats

Jonah Bellemans*[0009−0006−4336−6370], Qianying Liao*[0000−0002−5333−3103], Laurens Sion[0000−0002−8126−4491], Lieven Desmet[0000−0001−5155−7472], and Wouter Joosen[0000−0002−7710−5092]

DistriNet Research Group, KU Leuven
Celestijnenlaan 200a, 3001 Heverlee, Belgium
https://distrinet.cs.kuleuven.be/

Abstract. Generative Artificial Intelligence (GenAI) has rapidly evolved from an experimental technology into a foundational component of modern software systems. However, as its adoption grows, protecting sensitive personal data becomes increasingly challenging. Specifically, GenAI systems not only amplify traditional privacy threats but also introduce new inference-based risks, such as constructing detailed user profiles from seemingly harmless inputs. In response, privacy threat modeling frameworks are beginning to capture GenAI-specific privacy threats at a finer level of granularity. At the same time, a growing number of mitigation techniques have been proposed to address these threats. However, although knowledge of both threats and mitigations continues to mature, the problem- and solution-space have developed largely independently. This position paper argues that the primary challenge in GenAI privacy engineering is not the lack of knowledge about privacy threats or mitigation techniques, but the missing bridge between them. We decompose this gap into three sub-problems: (i) lack of fine-grained threat-to-mitigation mapping for GenAI systems, (ii) inapplicable solution-space assumptions in the GenAI context, and (iii) prioritization difficulty under GenAI constraints. We derive four recommendations for future mitigation-selection approaches, and outline a suggested approach that extends established threat-to-mitigation mapping methods to GenAI-specific threat characteristics. We propose a research agenda toward more systematic privacy mitigation selection for GenAI-based systems.

Keywords: GenAI · large language models · privacy threat modeling · privacy engineering · privacy by design · mitigation strategies

1 Introduction

Generative Artificial Intelligence (GenAI) systems are increasingly embedded in contemporary software systems, reshaping how data is processed, inferred, and exposed across a wide range of application domains. Compared with earlier

* These authors contributed equally to this work.

generations of AI systems, which were often developed for specialized tasks and deployed in tightly controlled environments, GenAI technologies significantly lower the barrier to adoption and integration. Through widely accessible APIs, software development kits, and foundation model services, developers can rapidly incorporate GenAI capabilities into existing applications with minimal machine learning expertise.

While the widespread accessibility of GenAI technologies has accelerated their adoption across the software ecosystem, it has also propagated privacy and identity-related risks into a broader range of applications and deployment contexts. A recent study found that "GPT-4 and ChatGPT reveal private information in contexts that humans would not, 39% and 57% of the time, respectively" [25]. Such findings raise significant concerns regarding the privacy implications of GenAI systems, particularly in enterprise environments processing identity-linked and organizational data. These risks span the entire GenAI lifecycle: models memorize and regurgitate training data, prompts and retrieval pipelines leak identity-linked information, and models infer sensitive attributes from seemingly innocuous inputs [8, 9, 35].

Furthermore, they have already materialized in everyday applications and enterprise systems. For instance, studies show that employees frequently submit sensitive corporate information to GenAI systems [27], with nearly 10% of enterprise GenAI prompts containing confidential data such as source code, financial records, or customer information [30]. In addition, GenAI systems have generated fabricated, privacy-invasive outputs involving real individuals, including a widely reported case in which ChatGPT falsely accused a law professor of sexual harassment [36].

Moreover, the growing body of research on GenAI privacy has led to increasingly fine-grained methods for eliciting, characterizing, and cataloging privacy threat scenarios [8, 12, 23, 24]. Among these efforts, LINDDUN4GenAI [23] extends the LINDDUN privacy threat modeling framework with GenAI-specific threat characteristics and Common Attacker Models (CAMs), significantly enriching the *problem-space* knowledge available to software engineers.

The *solution-space* has evolved just as rapidly. Established resources, including privacy design strategies and tactics [11, 18, 19], privacy patterns [10], and privacy-enhancing technologies (PETs) [17], have been complemented by a growing body of GenAI-specific mitigation techniques [31].

Although both the problem-space and solution-space continue to mature, a more fundamental challenge is that they have largely evolved independently. Consequently, the guidance linking identified threats to appropriate mitigations remains predominantly limited to high-level LINDDUN threat types, rather than leveraging the fine-grained threat characteristics captured during problem-space analysis. As demonstrated by Al-Momani et al. [2] for traditional LINDDUN, much of the threat detail obtained during problem-space analysis is effectively *lost in translation* when transitioning to the solution-space. This disconnect is even more pronounced in the GenAI domain, where many newly identified threat characteristics lack targeted mitigation guidance altogether. Furthermore,

even when multiple candidate mitigations are applicable, there is no principled basis for prioritizing among them under GenAI-specific privacy requirements and constraints. Therefore, software engineers lack actionable and systematic support for mapping GenAI-specific privacy threats to appropriate mitigations and for selecting among alternative mitigation strategies in a principled manner.

We argue that GenAI privacy engineering lacks neither threat knowledge nor mitigation techniques, but rather the bridge between them: as long as both continue to grow as separate bodies of knowledge, and matching the right mitigation to an identified threat remains unsupported by concrete, fine-grained selection methodologies, ever-finer threat modeling will not yield more private systems.

Addressing this gap requires GenAI-specific mitigation-selection approaches that do not yet exist. Rather than prescribing one, this paper formulates four recommendations that such approaches should satisfy, and suggests one potential approach: extending the more general methodology of Al-Momani et al. [2] to the GenAI-specific threat characteristics introduced by LINDDUN4GenAI [23]. Where relevant, such guidance should also be informed by legal obligations under frameworks such as the EU AI Act [16], so that selected mitigations can be situated within applicable regulatory expectations and data subject rights. In doing so, the paper outlines a research agenda for bridging the problem- and solution-space of GenAI privacy engineering.

The remainder of this paper is structured as follows. Section 2 provides background and discusses related work on privacy threat modeling, GenAI system paradigms, LINDDUN4GenAI and its threat characteristics, the layers of the solution-space: privacy design strategies and tactics, privacy patterns, and PETs, and finally existing efforts to bridge the gap between the problem- and the solution-space. Section 3 elaborates the three core sub-problems. Section 4 formulates recommendations for future mitigation-selection approaches and sketches a candidate direction extending the approach of Al-Momani et al. [2] to the GenAI context. Section 5 concludes with a call to action for the community.

2 Background and Related Work

This section introduces the concepts the remainder of the paper builds on: the practice of privacy threat modeling and the LINDDUN framework (section 2.1), its GenAI-specific extension LINDDUN4GenAI and the threat characteristics central to our argument (section 2.2), the layered solution-space of strategies, tactics, patterns, and PETs from which mitigations are drawn (section 2.3), and finally the available guidance to bridge the gap between threats and their suitable mitigations (section 2.4).

2.1 Privacy threat modeling

Threat modeling is a systematic approach to identifying what can go wrong in a system before it is built [33]. Originating in security engineering, approaches such as STRIDE [20] enumerate threat categories (or '*types*') over a model of

the system under design, and the method has since grown into a broad family of methods [38]. Because privacy harms are not reducible to security violations, dedicated privacy threat modeling frameworks have emerged.

Among these frameworks, LINDDUN [14, 15] is one of the most widely used. Starting from a model of the system under design, typically a Data Flow Diagram (DFD), it guides analysts through the elicitation of privacy threats in the seven threat types of its mnemonic: Linking, Identifying, Non-repudiation, Detecting, Data Disclosure, Unawareness & Unintervenability, and Non-compliance. The first five are commonly referred to as *hard* privacy threats, concerning properties such as anonymity, unlinkability, and confidentiality; the latter two as *soft* privacy threats, concerning transparency, intervenability, and compliance [14]. For each type, LINDDUN provides a threat tree that refines the threat type into *threat characteristics*: the lower the node, the more specific the description of how, and due to which cause, the threat arises. This information is stored in an actively maintained and reusable threat knowledge base [34].

Complementary frameworks approach privacy threats from other angles. MITRE's PANOPTIC [32] offers an alternative privacy threat model rooted in a taxonomy of privacy threat activities; PRIAM [13] complements threat elicitation with privacy *risk* analysis, quantifying the severity of identified risks; and PLOT4AI [7] provides a practical threat library specifically for AI-based systems. These efforts differ in elicitation style and scope, but share LINDDUN's focus on the problem-space.

2.2 GenAI-based systems and LINDDUN4GenAI

Users and organizations interact with GenAI in four main deployment paradigms: (i) direct use of a pre-trained foundation model, (ii) interaction with a fine-tuned model adapted to a domain, (iii) use of an application built around a model through system prompts, and (iv) agentic systems in which the model is embedded in a broader software ecosystem and invokes external tools and data sources [23]. The paradigms differ in who controls the training data, the model weights, and the prompts, a distinction that becomes central to mitigation feasibility later in this paper.

LINDDUN4GenAI [23] extends the LINDDUN threat knowledge base to these systems. From a synthesis of the academic literature and two industrial case studies (a GenAI-based chatbot and an agentic assistant), it identifies six Common Attacker Models (CAMs): recurring types of leakage defined by the targeted information and the acting party. CAM1 and CAM2 concern user-to-system and system-to-user information leakage, while CAM3 covers leakage of information from the pre-training to the fine-tuning party. CAM4 and CAM5 concern leakage between the system and agents (in either direction). Finally, in CAM6 (*residual privacy leakage*), a party with access to the model's intermediate computations, such as embeddings, activations, or gradients, can invert them to recover personal data that was never disclosed in any output [23].

Based on these CAMs, LINDDUN4GenAI adds nine new GenAI-specific threat characteristics to the threat trees. Several of these are central to this paper.

Among the hard privacy characteristics, *DD.1.3 Data type structure* captures disclosure through the structure of data representations themselves; its child *DD.1.3.2 Data derivations* covers the model-internal derivations exploited in CAM6. *DD.3.5 Fabrication* captures the disclosure of incorrect, hallucinated personal information that is presented as if it were accurate.

The existing soft privacy characteristics *U.2.2 Access* and *U.2.3 Rectification/erasure* are refined with AI-specific children reflecting that hallucinated personal data leaves no records that could be accessed, rectified, or erased, and that data absorbed into model weights cannot, in general, be selectively untrained. Finally, *U.3 Interference with personal decision making* introduces an entirely new class of threat: the system interfering with the autonomy of the data subject, for example through manipulation. Two further characteristics concern non-compliance with the EU AI Act (Nc.1.3) and with AI standards and best practices (Nc.4.2).

2.3 Privacy mitigations: strategies, tactics, patterns, and PETs

The solution-space knowledge available to privacy engineers is organized into four layers of decreasing abstraction, as depicted in fig. 1.

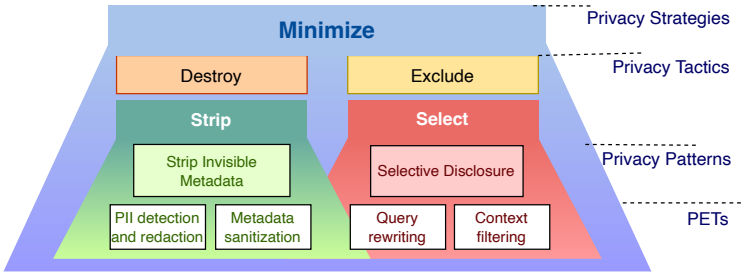

Fig. 1. Hierarchy of privacy solution-space knowledge illustrated using the *Minimize* privacy strategy as an example. In this example, the *Minimize* strategy is refined into the tactics *Strip* and *Select*. The *Strip* tactic is exemplified by the *Strip Invisible Metadata* pattern, which may be implemented using PETs such as PII (personally identifiable information) detection and redaction or metadata sanitization. Likewise, the *Select* tactic is illustrated by the *Selective Disclosure* pattern, which may be realized using PETs such as query rewriting and context filtering (note: the listed PETs here are merely illustrative, they do not form a complete mapping). The hierarchy demonstrates the increasing level of implementation specificity from strategies to PETs and motivates the solution-space organization.

At the top, *privacy design strategies* express architectural goals: Hoepman [18, 19] distinguishes four data-oriented strategies (minimize, separate, abstract[1], hide) and four process-oriented ones (inform, control, enforce, demonstrate).

For the second layer, Colesky et al. [11] refine each strategy into *tactics*, a refinement later consolidated in Hoepman's *little blue book* [19]; the hide strategy, for instance, comprises *restrict*, *mix*, *obfuscate*, and *dissociate* [11, 19].

At the third layer, *privacy patterns* describe reusable solutions to recurring design problems and are compiled in a community catalog [10], classified by the strategies and tactics they realize.

At the bottom, *privacy-enhancing technologies* (PETs) are concrete techniques implementing (parts of) patterns, ranging from attribute removal to differential privacy and homomorphic encryption, and have been organized in dedicated taxonomies [17]. Moving down the layers, guidance becomes more concrete but also more context-dependent: which PET is appropriate depends on properties of both the system and the threat, which is where the selection problem highlighted by this paper arises.

2.4 Mitigation selection guidance

A first line of work supports the selection of mitigations without starting from elicited privacy threats. Kunz and Binder [22] observe that existing PET systematizations are outdated or focus on comparison criteria rather than on guidance for practical selection, and propose a classification that adds functional context, technology maturity, and impact on non-functional requirements to existing criteria, such as the privacy protection goal. Pape et al. [29] anchor mitigation selection in data protection law instead, mapping trust models and PET types onto GDPR principles. Baldassarre et al. [6] formalize the relations among privacy-by-design principles, design strategies, patterns, and vulnerabilities in a privacy knowledge base with accompanying tool support. The entry point of these approaches is the functional context of the system, a legal principle, or a vulnerability, rather than a privacy threat identified during analysis. The fine-grained, specific threat information obtained during threat elicitation therefore plays no role in the selection of suitable mitigations with these techniques.

A selection methodology that does take the elicited threat information into account, was proposed by the authors of LINDDUN itself in its earlier stages [14]. Deng et al. accompany the threat trees with a taxonomy of mitigation strategies and a classification of privacy-enhancing solutions, linking each threat type to a mitigation strategy and, through it, to candidate PETs [14]. However, the mapping operates at the level of threat types, not the more fine-grained threat characteristics. For example, it indicates what to do about identifying threats in general, but not what to do about specific causes of identifying threats elicited within the system. Furthermore, the LINDDUN threat knowledge base has since

[1] Hoepman initially named this strategy 'aggregate' [18], but it was later renamed by Colesky et al. [11]. The new naming is also reflected in Hoepman's little blue book [19].

been reworked and restructured multiple times [34], as well as extended to GenAI [23], without a corresponding update on the mitigation side.

On the GenAI side, the closest counterpart to such a mapping is a survey by Wang et al. [37]. It divides the LLM life cycle into four scenarios, namely pre-training, fine-tuning, deployment, and LLM-based agents, which parallel the deployment paradigms mentioned by LINDDUN4GenAI [23], pairing the risks of each scenario with countermeasures. Their comparison tables record, for each countermeasure, the risk it aims to mitigate, the access the defender would require to the model and the training data, the models and tasks to which it applies, and its reported limitations. However, the countermeasures are once again mapped to higher-level risks, rather than specific threat characteristics attached to a model of the system under design. Other SoK papers follow the same pattern: they pair risks with mitigations for a single paradigm or a single class of techniques, and assess maturity rather than fit to an identified threat [8, 31].

To our knowledge, the only works that bridge fine-grained classical LINDDUN threats to concrete mitigations are those of Al-Momani et al.[2] [2–5]. First, they analyze 70 patterns of a widely used privacy pattern catalog along five property dimensions that govern selection (applicability scope, privacy objective, impacted qualities, data focus, and hotspot), and find that these properties are rarely made explicit in the catalog itself [3]. Subsequently, they show that the fine-grained knowledge gained during threat analysis is *lost in translation* from the problem-space to the solution-space, and propose a remedy: 36 *key nodes* identified in the LINDDUN threat trees, grouped into ten mitigation goals, with solution flowcharts for the four goals that admit many candidate countermeasures [2]. In later experiments, they evaluate and build further upon this work in practice. Eliciting threats for a robotaxi service with LINDDUN and matching them against the pattern catalog, they report inconsisties within the patters, a lack of guidance on applying these patterns, as well as threats for which no suitable pattern exists [5]. Applying three PET-selection approaches (including their own) to the same setting, they conclude that none adequately supports selection in a realistic scenario, and derive requirements for future selection methodologies [4].

3 Problem Statement

The gap this paper addresses is best characterized by highlighting where the current state of the art falls short.

On the problem-space side, LINDDUN4GenAI [23] extends the LINDDUN threat trees with GenAI-specific threat characteristics and Common Attacker Models, but deliberately stops at threat identification: it prescribes neither which mitigation addresses a given threat, nor how to choose among alternatives. As mentioned in section 2.4, the existing threat-to-mitigation mapping of LINDDUN [14] is insufficiently fine-grained, has not been updated to reflect the latest

[2] While these works all share the same first author, we note that the co-authors involved only partially overlap and may differ per publication.

knowledge base contents and structure, and predates the GenAI threat characteristics. Adjacent security-oriented resources do not fill this gap either. The OWASP Top 10 for LLM Applications [28] and MITRE ATLAS [26] index mitigations by vulnerability and adversarial technique, not by privacy threat.

On the solution-space side, we examine the line of work of Al-Momani et al. [2–5] in more detail, since it is the only one connecting fine-grained LINDDUN threat knowledge to concrete mitigations, and since it reports on applying that connection to a real-world case.

Even for classical LINDDUN, their results leave the solution-space unevenly covered, in two complementary ways. On the side of the pattern catalogs, they show that the available patterns predominantly serve soft privacy goals such as transparency and intervenability, while far fewer address hard privacy goals such as anonymity and unlinkability [3]. In later work, they further quantify this imbalance, reporting 59 process-oriented patterns versus 19 data-oriented patterns[3] [5]. On the guidance side, their decision support [2] covers only the five hard privacy threat types and explicitly excludes the soft ones, on the grounds that selection support for soft privacy countermeasures has been treated elsewhere in the form of pattern systems. **Each group therefore lacks what the other has: structured selection guidance for hard privacy threats, but few patterns to select from; many patterns for soft privacy threats, but guidance organized by design strategy rather than by elicited threat.** Their robotaxi case study [5] indicates that this leaves engineers without a usable route from a threat to a pattern. They report inconsistent levels of abstraction across patterns and a lack of guidance for finding the right pattern for a given type of privacy threat. In addition, they had to propose new patterns for threats the catalog does not address at all, in particular for data deletion and for replacing sensitive data [5].

The two most recent papers also evaluate their methodology in practice, and report that it still has shortcomings which need to be addressed. Matching each elicited threat of the robotaxi design against the pattern catalog, **they find that mitigating a single threat frequently requires a combination of patterns for which no composition guidance exists, and that even once a suitable pattern has been identified, selecting a concrete PET to realize it remains difficult [5].** The most recent study applies three published PET-selection approaches [2, 21, 22], complemented by a pragmatic approach based on privacy design strategies [18], to the same use case, and concludes that none of them adequately supports PET selection in a complex real-world scenario [4]. They make three observations relevant for the argument of this paper. First, the approaches yield a set of applicable PETs but offer little support for the final choice among them. Second, each approach treats threats in isolation and selects at least one PET per threat, whereas both threats and PETs are interdependent in practice. Third, the solution flowcharts of Al-Momani et al. [2] prove sparsely populated: the chart for protecting identity lists attribute-based

[3] In this work, Al-Momani et al. report more than 70 patterns, indicating that the pattern catalog has grown since their "land of the lost" paper in 2021 [3].

credentials as the only countermeasure, and the chart for protecting attributes refers to encryption in general without naming a specific technology.

Finally, the line of work of Al-Momani et al. predates the GenAI-specific threat knowledge base of LINDDUN4GenAI. The pattern properties of Al-Momani et al. [3] are defined for traditional software architectures, and their robotaxi studies concern a cyber-physical system in which machine learning is a component of the use case rather than the object of the privacy analysis [4, 5]. Their key nodes and flowcharts [2] rest on an earlier generation of the LINDDUN threat trees. The structure and node identifiers of that generation differ from the restructured knowledge base on which LINDDUN4GenAI builds. This issue was explicitly mentioned in their follow-up case study, in which they mention that their approach had to be modified because it had been designed for this earlier version of LINDDUN, rendering the use of the key nodes unfeasible [4]. **None of these results covers any of the GenAI-specific threat characteristics. They are thus not tailored to the problem- and solution-space that GenAI introduces.**

We decompose the resulting gap into three sub-problems.

Sub-problem 1: Lack of fine-grained threat-to-mitigation mapping for GenAI systems

Existing threat-to-mitigation guidance operates at the level of top-level threat types or high-level design strategies [11, 18, 19], discarding the fine-grained knowledge encoded in the lower nodes of the threat trees. Al-Momani et al. [2] provide finer-grained threat-to-mitigation guidance for classical LINDDUN by identifying *key nodes* at which candidate mitigations diverge. The GenAI threat characteristics, however, did not exist when their flowcharts were constructed, so no key nodes are defined for them.

For example, consider the LINDDUN4GenAI threat characteristic *DD.1.3.2 Data derivations*, where intermediate representations such as embeddings leak personal data, allowing otherwise pseudonymous users to be re-identified. Even if the analyst selects a fitting strategy for their situation, the strategy label alone cannot discriminate among the different mitigations it subsumes. For example, in the *hide* strategy, (i) restricting access to the embedding store, (ii) encrypting stored vector databases, and (iii) confining inference to a trusted execution environment could all potentially fit into the existing guidance [2], which lists access control, encryption, or secure processing as potential countermeasures. However, which of these countermeasures apply depends on the cause of the leakage, namely whether it arises from memorization during training, exposure of stored representations, or embeddings exchanged with third parties.

The information needed to determine which option is most suitable resides in the leaf nodes of the threat tree, which the category-level guidance discards. Nor can the existing flowcharts be reused: their applicability questions ("are all used attributes required for the system?", "could sensitive attributes be replaced with different, yet less sensitive, attributes?" [2]) presume attributes an engineer

deliberately chose to collect, and are ill-posed for representations that arise as byproducts of model architecture and training.

Sub-problem 2: Inapplicable solution-space assumptions in the GenAI context

Even where the solution-space offers building blocks, their coverage and their operating assumptions do not match the GenAI threat landscape.

First, the coverage asymmetry outlined above works against GenAI: the threat literature synthesized in LINDDUN4GenAI is dominated by leakage threats (memorization, extraction, and inference of training, fine-tuning, and prompt data) that map to hard privacy threat types [23], where the pattern supply is thinnest [3].

Second, the soft privacy patterns that do exist rest on an assumption that GenAI invalidates: that the system holds identifiable stored records that a data subject can access, rectify, or erase. LINDDUN4GenAI documents that this assumption fails twice over: for hallucinated personal data "there are simply no records to 'rectify' or 'delete' ", as (i) the system is inherently probabilistic and may fabricate a different claim on each query (U.2.2.1, U.2.3.1), and (ii) personal data absorbed into model weights cannot, in general, be selectively untrained (U.2.3.2) [23]. A privacy dashboard pattern applied to a system that fabricates a defamatory biography on the fly (DD.3.5) satisfies the letter of the pattern but misses the threat entirely.

Third, the pattern properties cataloged for selection support [3] omit properties that actually govern applicability in GenAI systems, foremost whether the deploying party controls the training data, the model weights, or only the prompts. A GenAI threat-to-mitigation mapping must therefore do more than assign existing mitigations to new threats: it must annotate them with GenAI-relevant applicability properties, and make explicit where no adequate mitigation currently exists. Such GenAI-relevant properties can be found in more recent GenAI-mitigation research, such as the work of Kunz and Binder [22], or Wang et al. [37].

Sub-problem 3: Prioritization difficulty under GenAI constraints

When several mitigations are applicable, engineers need a principled ordering. The only ordering available in the fine-grained bridge is the one Al-Momani et al. [2] adopt in their flowcharts: countermeasures of the *minimize* strategy first, then *separate*, *abstract*, and *hide*, so that options with little utility impact are examined before options carrying significant computational or communication overhead. The strategies themselves come with no such ordering [18, 19], and Al-Momani et al. present their ordering as just one example a designer of selection support might adopt [2]. Furthermore, their ordering presumes that the threat concerns *collected* data flowing through the system: each strategy reduces the amount, linkability, or precision of data already in flow. It offers no guidance for threats concerning *generated* content or *influence on the data subject*. For

generated content, minimization presupposes knowing what personal data is in flow, whereas neither the form in which that data will appear nor its correctness is known in advance. For influence on the data subject, there is no data to minimize in the first place.

DD.3.5 Fabrication concerns personal data the system invents rather than collects; *U.3 Interference with personal decision making*, a soft privacy threat, concerns the model's effect on the subject's decision-making rather than any data asset. Neither is covered by the existing ordering, and Al-Momani et al. [2] explicitly scope out the soft privacy threat type in which U.3 resides.

Moreover, another GenAI-specific factor compounds the difficulty: the deployment paradigm (pre-trained, fine-tuned, prompt-based, or agentic [23]) determines which mitigations are feasible at all. A technique such as Differentially Private Stochastic Gradient Descent (DP-SGD) [1], for instance, cannot be applied to a model consumed through a third-party API, since the consumer never controls training. While existing work such as that of Wang et al. [37] organize their GenAI risks and countermeasures based on this deployment paradigm, they do not use it to filter infeasible mitigations or to order the remaining ones.

4 Toward GenAI-Aware Mitigation Selection

The sub-problems of section 3 delineate what any future mitigation-selection approach for GenAI privacy threats must accomplish. In this section, we first formulate the recommendations that candidate approaches should follow (section 4.1), and then sketch one concrete direction that we consider promising (section 4.2), to illustrate that these recommendations are attainable.

4.1 Recommendations for future approaches

We derive recommendations from the sub-problems identified in section 3: R1 addresses sub-problem 1, R2 addresses sub-problem 2, and R3 addresses sub-problem 3, while R4 is a cross-cutting adoption concern. These recommendations are deliberately framework-agnostic: although this paper uses LINDDUN4GenAI as its anchor, they should hold for any approach that bridges a GenAI privacy threat knowledge base and a body of mitigations.

R1: Fine-grained threat-to-mitigation guidance. Selection guidance must operate at the granularity at which threats are identified: at the level of threat causes rather than broad threat types. Whatever the underlying threat knowledge base, be it the LINDDUN4GenAI threat trees [23] or other GenAI threat taxonomies [24, 31, 37], the fine-grained knowledge it captures must be used for mitigation selection rather than dismissed. Relying on key nodes [2] is one existing instantiation of this principle (sub-problem 1).

R2: Explicit applicability properties. Candidate mitigations must be annotated with the properties that actually determine their applicability in GenAI systems, e.g., (i) whether the deploying party controls the training data,

the model weights, or only the prompts; (ii) the impact on model utility; and (iii) the assumptions the mitigation relies on, such as the existence of identifiable, rectifiable records. Where no adequate mitigation exists at all, this absence must be made explicit (sub-problem 2).

R3: GenAI-valid prioritization principles. Ordering principles must hold in the GenAI context: (i) feasibility filtering by deployment paradigm before any ranking; (ii) a cost model reflecting GenAI-specific overhead; and (iii) dedicated treatment of threats concerning generated content and influence on the data subject, for which minimization-first logic does not apply (sub-problem 3).

R4: Compatibility and evolvability. To be adopted, the guidance should integrate with established privacy engineering and threat modeling practice rather than require a parallel process, and it should remain maintainable as models, deployment patterns, and attacks evolve, for instance through an extensible, publicly maintained knowledge base as exemplified by LINDDUN4GenAI [23].

4.2 A candidate direction: extending the key-node method

The construction approach of Al-Momani et al. [2] is generic by design: it prescribes how to build selection support from a threat knowledge base. One promising direction is therefore to re-instantiate this approach on the LINDDUN4GenAI threat trees, with GenAI-specific modifications at each step.

First, key nodes would be identified for the GenAI-specific threat characteristics (R1). The lower nodes of the new trees carry the cause information that key-node identification requires. The children of *DD.1.3 Data type structure*, for instance, distinguish leakage through the data encoding itself from leakage through model-internal derivations such as embeddings and gradients [23].

Second, the resulting key nodes would be grouped into mitigation goals. We anticipate that some goals will have no classical counterpart, such as preventing or containing fabricated personal data (DD.3.5) and constraining the system's influence on the data subject (U.3), and that grouping may need to span CAMs rather than DFD elements alone.

Third, candidate mitigations would be assembled for each goal from the existing solution-space layers (strategies and tactics [11, 18, 19], patterns [10], PETs [17]) and from GenAI-specific techniques cataloged in recent surveys [31], each annotated with the applicability properties and explicit gap flags where no adequate option exists (R2).

Fourth, the ordering step would be revised (R3). For hard privacy threats concerning collected data, the utility-loss rationale behind the existing minimize-first ordering may be retained, though recalibrated to GenAI cost structures; homomorphic encryption, for example, is even less feasible for Large Language Model (LLM) inference than for classical processing. Before any ranking, a feasibility filter based on the deployment paradigm would discard mitigations the deploying party cannot apply. For generated content and influence threats (DD.3.5, U.3), dedicated orderings would be needed, grounded in the deployment

paradigm and in the nature of the threat actor, whether a user, the system itself, or an agent.

Finally, applicability questions and output representations analogous to the existing solution flowcharts [2] would keep the result recognizable to analysts already familiar with that method (R4).

We emphasize that this is a direction, not a result: constructing the mapping, defining the heuristics, and validating both with practitioners remain open research tasks.

5 Conclusion

Privacy threat modeling for GenAI-based systems and the development of GenAI privacy mitigations are both advancing, but as separate, still-maturing bodies of knowledge: matching the right mitigation to an identified threat remains unsupported. LINDDUN4GenAI [23] provides fine-grained, GenAI-specific problem-space knowledge, yet (i) there is no fine-grained guidance that connects GenAI-specific threat knowledge to fitting mitigations; (ii) traditional solution-space assumptions do not hold up in the GenAI context; and (iii) there is no ordering principle valid under GenAI constraints.

We argue that closing this gap is a research task for the privacy engineering community, not an exercise in applying existing methods. To that end, we formulated four recommendations that future mitigation-selection approaches should satisfy and sketched one candidate direction: extending the key-node method of Al-Momani et al. [2] to the LINDDUN4GenAI threat trees, with mitigation goals, applicability properties, and orderings revised for GenAI. Elaborating this direction, constructing the actual threat-to-mitigation mapping, and validating it in practice fall outside the scope of a position paper and are left as future work.

Closing the gap will also require expertise beyond computer science. Data protection principles such as storage limitation and minimization must be interpreted in ways that are technically meaningful for models that memorize their training data, and regulatory expectations under frameworks such as the EU AI Act [16] must be connectable to concrete mitigations. Without such interdisciplinary work, data subject rights risk remaining formal: acknowledged in policy, but not enforced at the level of system design. By articulating this absence of mitigation-selection guidance for GenAI privacy threats, we aim to motivate research that bridges the gap.

Acknowledgments. This research is partially funded by the Research Fund KU Leuven, Internal Funds KU Leuven, and by the Cybersecurity Research Program Flanders.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

GenAI Statement GenAI-based tools were used to revise the text, improve flow, and correct any typos, grammatical errors, and awkward phrasing.

References

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep Learning with Differential Privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 308–318. CCS '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2976749.2978318, https://doi.org/10.1145/2976749.2978318

[2] Al-Momani, A., Bösch, C., Wuyts, K., Sion, L., Joosen, W., Kargl, F.: Mitigation Lost in Translation: Leveraging Threat Information to Improve Privacy Solution Selection. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22). pp. 1236–1245. ACM, New York, NY, USA (2022). https://doi.org/10.1145/3477314.3507107

[3] Al-Momani, A., Wuyts, K., Sion, L., Kargl, F., Joosen, W., Erb, B., Bösch, C.: Land of the Lost: Privacy Patterns' Forgotten Properties: Enhancing Selection-Support for Privacy Patterns. In: Proceedings of the 36th ACM/SIGAPP Symposium on Applied Computing (SAC '21). pp. 1217–1225. ACM, New York, NY, USA (2021). https://doi.org/10.1145/3412841.3441996

[4] Al-Momani, A., Balenson, D., Bösch, C., Mann, Z.Á., Pape, S., Petit, J.: Lessons from a Robotaxi: Challenges in Selecting Privacy-Enhancing Technologies, pp. 154–170. Springer Nature Switzerland (2026). https://doi.org/10.1007/978-3-032-16089-8_11

[5] Al-Momani, A., Balenson, D., Mann, Z.Á., Pape, S., Petit, J., Bösch, C.: Navigating Privacy Patterns in the Era of Robotaxis. In: 2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). pp. 32–39. IEEE (Jul 2024). https://doi.org/10.1109/eurospw61312.2024.00011

[6] Baldassarre, M.T., Barletta, V.S., Caivano, D., Piccinno, A., Scalera, M.: Privacy Knowledge Base for Supporting Decision-Making in Software Development, pp. 147–157. Springer International Publishing (2022). https://doi.org/10.1007/978-3-030-98388-8_14

[7] Barberá, I.: PLOT4AI: Practical Library of Threats for Artificial Intelligence. https://plot4.ai/ (2025), accessed: 2025-11-30

[8] Bodea, A.E., Meisenbacher, S., Klymenko, A., Matthes, F.: Sok: Privacy risks and mitigations in retrieval-augmented generation systems. arXiv preprint arXiv:2601.03979 (2026)

[9] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al.: Extracting training data from large language models. In: 30th USENIX security symposium (USENIX Security 21). pp. 2633–2650 (2021)

[10] Colesky, M., Hoepman, J.H., Boesch, C., Kargl, F., Kopp, H., Mosby, P., Métayer, D.L., Drozd, O., del Álamo, J.M., Martin, Y.S., Caiza, J.C., Gupta, M., Doty, N.: Privacy Patterns. https://privacypatterns.org/patterns/, Last Accessed: 2026-07-27

[11] Colesky, M., Hoepman, J.H., Hillen, C.: A critical analysis of privacy design strategies. In: 2016 IEEE Security and Privacy Workshops (SPW). pp. 33–40. IEEE (2016). https://doi.org/10.1109/SPW.2016.23

[12] Das, B.C., Amini, M.H., Wu, Y.: Security and Privacy Challenges of Large Language Models: A Survey. ACM Computing Surveys **57**(6), 1–39 (Feb 2025). https://doi.org/10.1145/3712001

[13] De, S.J., Le Métayer, D.: PRIAM: A privacy risk analysis methodology. Lecture Notes in Computer Science pp. 221–229 (2016)

[14] Deng, M., Wuyts, K., Scandariato, R., Preneel, B., Joosen, W.: A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. Requirements Engineering **16**(1), 3–32 (2011)
[15] DistriNet, KU Leuven: LINDDUN Website. https://linddun.org/pro (2025)
[16] European Union: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the EU **L**(2024/1689), 1–144 (Jun 2024), http://data.europa.eu/eli/reg/2024/1689/oj
[17] Heurix, J., Zimmermann, P., Neubauer, T., Fenz, S.: A taxonomy for privacy enhancing technologies. Computers & Security **53**, 1–17 (2015). https://doi.org/10.1016/j.cose.2015.05.002
[18] Hoepman, J.H.: Privacy design strategies. In: Cuppens-Boulahia, N., Cuppens, F., Jajodia, S., Abou El Kalam, A., Sans, T. (eds.) ICT Systems Security and Privacy Protection – IFIP TC 11 29th International Conference (SEC 2014). pp. 446–459. Springer, Berlin, Heidelberg (2014). https://doi.org/10.1007/978-3-642-55415-5_38
[19] Hoepman, J.H.: Privacy Design Strategies (The Little Blue Book) (2022), https://cs.ru.nl/~jhh/publications/pds-booklet.pdf, version of April 19, 2022
[20] Kohnfelder, L., Garg, P.: The threats to our products. Microsoft Interface, Microsoft Corporation **33**, 1–8 (1999)
[21] Kunz, I., Banse, C., Stephanow, P.: Selecting Privacy Enhancing Technologies for IoT-Based Services, pp. 455–474. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-63095-9_29
[22] Kunz, I., Binder, A.: Application-Oriented Selection of Privacy Enhancing Technologies, pp. 75–87. Springer International Publishing (2022). https://doi.org/10.1007/978-3-031-07315-1_5
[23] Liao, Q., Bellemans, J., Sion, L., Jiang, X., Usynin, D., Zhou, X., Van Landuyt, D., Desmet, L., Joosen, W.: AI've Got a Bad Feeling About This: A Privacy Threat Modeling Framework for GenAI. In: SOUPS '26: Proceedings of the Twenty-Second USENIX Conference on Usable Privacy and Security. USENIX Association, Hannover, Germany (Aug 2026)
[24] Ma, B., Jiang, Y., Wang, X., Yu, G., Wang, Q., Sun, C., Li, C., Qi, X., He, Y., Ni, W., et al.: Sok: Semantic privacy in large language models. arXiv preprint arXiv:2506.23603 (2025)
[25] Mireshghallah, N., Kim, H., Zhou, X., Tsvetkov, Y., Sap, M., Shokri, R., Choi, Y.: Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (eds.) International Conference on Learning Representations. vol. 2024, pp. 1892–1915 (2024), https://proceedings.iclr.cc/paper_files/paper/2024/file/08305d8b2ddab98932c163ea73df065f-Paper-Conference.pdf
[26] MITRE Corporation: MITRE ATLAS (2025), https://atlas.mitre.org/
[27] Muncaster, P.: One Fifth of CISOs Say Staff Leaked Sensitive Data Through GenAI. Infosecurity Magazine (Apr 2024), https://www.infosecurity-magazine.com/news/fifth-cisos-staff-leaked-data-genai/
[28] OWASP Foundation: OWASP Top 10 for Large Language Model Applications version 1.1 (2024), https://owasp.org/www-project-top-10-for-large-language-model-applications/

[29] Pape, S., Bkakria, A., Chah, B., Heymann, M., Syed-Winkler, S.: A Framework for Supporting PET Selection Based on GDPR Principles, pp. 3–23. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-032-00624-0_1
[30] Schuman, E.: Nearly 10% of Employee Gen AI Prompts Include Sensitive Data. CSO Online (Feb 2025), https://www.csoonline.com/article/3819170/nearly-10-of-employee-gen-ai-prompts-include-sensitive-data.html
[31] Shanmugarasa, Y., Ding, M., Arachchige, C.M., Rakotoarivelo, T.: SoK: The privacy paradox of large language models: Advancements, privacy risks, and mitigation. In: Proceedings of the 20th ACM Asia Conference on Computer and Communications Security. pp. 425–441 (2025)
[32] Shapiro, S., Bloom, C., Ballard, B., Slotter, S., Paes, M., McEwen, J., Xu, R., Katcher, S.: The PANOPTIC™ Privacy Threat Model. Tech. Rep. Version 1, MITRE (Dec 2023)
[33] Shostack, A.: Threat Modeling: Designing for Security. John Wiley & Sons, Indianapolis, Indiana (2014)
[34] Sion, L., Van Landuyt, D., Wuyts, K., Joosen, W.: Robust and reusable LINDDUN privacy threat knowledge. Computers & Security **154**, 104419 (Jul 2025). https://doi.org/10.1016/j.cose.2025.104419
[35] Staab, R., Vero, M., Balunovic, M., Vechev, M.: Beyond Memorization: Violating Privacy via Inference with Large Language Models. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (eds.) International Conference on Learning Representations. vol. 2024, pp. 33832–33878 (2024), https://proceedings.iclr.cc/paper_files/paper/2024/file/9028b8a3ca98f58e373f0c1497a17448-Paper-Conference.pdf
[36] Verma, P., Oremus, W.: ChatGPT invented a sexual harassment scandal and named a real law prof as the accused. https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/ (Apr 2023)
[37] Wang, S., Zhu, T., Liu, B., Ding, M., Ye, D., Zhou, W., Yu, P.: Unique Security and Privacy Threats of Large Language Models: A Comprehensive Survey. ACM Computing Surveys **58**(4), 1–36 (Oct 2025). https://doi.org/10.1145/3764113
[38] Xiong, W., Lagerström, R.: Threat modeling–A systematic literature review. Computers & Security **84**, 53–69 (2019)