LLM Guardrails for Privacy by Design: A Privacy-Threat Taxonomy and Black-Box Evaluation

Rui Zhang[0009−0007−5127−5185] and Dimitri Van Landuyt[0000−0001−6597−2271]

LIRIS Research Centre Information Systems Engineering,
Faculty of Economics and Business (FEB), KU Leuven, Belgium
rui.zhang1@kuleuven.be, dimitri.vanlanduyt@kuleuven.be

Abstract. Large language models increasingly process requests involving personal, sensitive, and context-dependent information. However, Privacy by Design principles, structured privacy-threat analysis, and the empirical evaluation of model-level guardrails remain insufficiently connected. This study adopts complementary top-down and bottom-up approaches. The top-down approach synthesises privacy-related literature to develop a six-category, test-oriented privacy-threat taxonomy and identify the corresponding Privacy by Design objectives. The bottom-up approach empirically evaluates privacy-related guardrail behaviour through black-box testing. The results provide evidence of privacy-related guardrails, while also revealing a substantial gap between the top-down privacy-threat taxonomy and the bottom-up categories represented in a general-purpose safety benchmark. Overall, the study provides a structured conceptual and empirical basis for privacy-oriented guardrail evaluation.

Keywords: Privacy by Design · LLM Guardrails · Privacy Threat Modelling · Black-box Testing

1 Introduction

Large language models increasingly process requests involving personal, sensitive, and context-dependent information. Such interactions may expose personal data, infer sensitive attributes from indirect cues, link information across sources, support profiling, or enable contextually inappropriate information use [22, 4, 14, 15]. These privacy threats extend beyond training-data memorisation and direct disclosure, creating a need for controls that can restrict inappropriate privacy-related requests at the model level.

Privacy by Design requires privacy protection to be embedded throughout system design and operation [3, 10, 12]. Although model-level guardrails cannot implement Privacy by Design on their own, they may serve as a complementary control by restricting requests involving the inappropriate disclosure, inference, linking, or use of personal information.

However, privacy-threat research and model guardrail evaluation remain insufficiently connected. Traditional privacy engineering and more recent research on privacy in LLMs provide relatively systematic accounts of privacy threats, but primarily support system-level threat modelling, risk governance, and the design of protective measures [8, 21, 15, 4]. Guardrail evaluation, by contrast, typically examines refusal behaviour, response safety, over-refusal, and robustness through externally observable prompt-response interactions [11, 20, 23]. However, these approaches are generally not organised around a structured, top-down conceptualisation or taxonomy of privacy threats. Consequently, it remains difficult to determine which privacy threats are covered by existing benchmarks and how their observed restriction behaviour relates to Privacy by Design objectives. Further work is therefore required to translate structured privacy threats into testable objects for prompt-response analysis and to combine multiple forms of black-box evidence when evaluating how model-level guardrails support privacy protection.

This study addresses this problem through complementary top-down and bottom-up paths. The top-down path synthesises established privacy-engineering frameworks and recent LLM privacy research to construct a six-category, test-oriented privacy-threat taxonomy and identify the corresponding Privacy by Design objectives. The bottom-up path uses privacy-related prompts extracted from WildGuardTest, a safety evaluation set released with WildGuard and organised according to its general-purpose safety taxonomy [11]. It applies three complementary testing strategies: a Chi-square comparison of response distributions between harmful and benign prompts, informed by contrastive safety evaluation [20]; diagnostic Self-report questions for distinguishing policy-based restriction from capability limitations [17]; and Shallow Depth testing based on assistant-prefill manipulation [23]. These strategies compare the externally observable behaviour of the original GPT-OSS 20B model with that of its abliterated variant. This design examines whether the model selectively restricts privacy-related requests and how these restrictions vary across testing conditions and model variants.

This study makes three contributions. First, it proposes a literature-derived, test-oriented privacy-threat taxonomy that connects observable privacy threats to Privacy by Design objectives. Second, it applies three complementary black-box testing strategies to privacy-related prompts and uses a contrast between the original and abliterated model variants to analyse the selectivity and robustness of model-level guardrails. Third, it compares this taxonomy with the privacy categories represented in WildGuardTest, revealing scope and coverage gaps between a general-purpose safety benchmark and privacy-specific guardrail evaluation.

2 Background

2.1 Privacy by Design and Privacy Threats

Privacy by Design requires privacy protection to be embedded in system architecture, default settings, and data-processing practices throughout the system lifecycle [3, 10]. Its objectives extend beyond preventing disclosure and include data minimisation, purpose limitation, access control, transparency, and appropriate information use [12].

This study defines a privacy threat as a request, condition, or potential model behaviour that may enable protected information to be inappropriately accessed, inferred, linked, used, or propagated. A privacy threat differs from a complete privacy-risk assessment, which additionally considers likelihood, context, affected parties, and potential impact [8, 21].

Because black-box prompt testing cannot reliably estimate all of these factors, the study develops a privacy-threat taxonomy rather than a complete privacy-risk assessment framework. Different threats also imply different Privacy by Design objectives, such as confidentiality, unlinkability, purpose limitation, and data minimisation.

2.2 Black-box Evaluation of Guardrail Behaviour

Black-box evaluation analyses model behaviour through observable inputs and outputs without relying on internal parameters, activations, system instructions, or safety classifiers. It can identify behavioural patterns, but cannot directly establish which internal mechanism produced a response.

Non-execution does not necessarily indicate guardrail activation. It may also result from missing knowledge, lack of access, misunderstanding, or limited capability. Similarly, a model may restrict sensitive content without issuing an explicit refusal. Refusal rates and isolated prompt-response outcomes are therefore insufficient for reliably interpreting guardrail behaviour [20, 19].

This study distinguishes four observable outcomes: guardrail-based restriction, incapacity-based non-execution, substantive compliance, and ambiguous behaviour when the available evidence is insufficient or conflicting. Because a single response cannot reliably distinguish these outcomes, the study uses multiple black-box testing strategies to obtain complementary evidence concerning behavioural selectivity, attribution, and robustness [17, 23].

3 Problem Statement

3.1 Motivation

Privacy-engineering research provides established frameworks for identifying threats such as information disclosure, identifiability, linkability, inappropriate information use, and excessive data processing [8, 21, 12, 18]. More recent LLM privacy

research extends this threat space to sensitive-attribute inference, generated personal information, and the reuse or propagation of information across interactions and contexts [22, 4, 14, 15].

However, these threat concepts have not been sufficiently operationalised for black-box guardrail evaluation. As a result, it remains difficult to determine which privacy threats are represented in existing safety benchmarks and how they relate to Privacy by Design objectives. Existing guardrail evaluations primarily examine general safety behaviour, refusal, over-refusal, and adversarial robustness, while privacy is typically represented as one broad category within a wider safety taxonomy [11, 20, 23].

Moreover, an isolated non-execution response cannot establish whether the observed behaviour results from a model-level guardrail, limited capability, missing information, or task misunderstanding [20, 17]. Evaluating the contribution of model-level guardrails to privacy protection therefore requires both a structured, test-oriented account of the privacy-threat space and complementary black-box evidence concerning behavioural selectivity, attribution, and robustness.

3.2 Research Questions

Based on this problem, the study addresses the following research questions:

RQ1: *How can privacy threats relevant to LLM-based systems be organised into a test-oriented taxonomy, and which Privacy by Design objectives do they implicate?*

RQ2: *To what extent do model-level guardrails exhibit observable privacy-related restriction behaviour under black-box testing, and what evidence do different testing strategies provide concerning its selectivity, attribution, and robustness?*

4 Methodology

This study combines a top-down conceptual analysis with a bottom-up empirical evaluation. The top-down path connects Privacy by Design principles, privacy threat modelling, privacy design strategies, concrete controls, and black-box evaluation. Table 1 summarises how these concepts are used in this study.

The bottom-up path evaluates prompts from WildGuardTest's high-level Privacy category using its three original subcategories. Because these categories cover only part of the broader top-down threat space, the experiment is not interpreted as a comprehensive evaluation of S1–S6; their correspondence is examined subsequently.

4.1 A Top-Down and Test-Oriented Privacy-Threat taxonomy

To address RQ1, we synthesise concepts and classifications from privacy threat modelling, privacy design strategies, AI privacy-risk taxonomies, and LLM privacy research [8, 21, 15, 6, 12, 14, 16, 4]. These bodies of work describe privacy

Table 1. Top-down analytical workflow for operationalising Privacy by Design in this study.

Analytical level	Concept or framework	Role in this study
Overarching approach	Privacy by Design and Data Protection by Design and by Default [3, 10]	Require privacy and data protection to be embedded into system architecture, default configurations, and processing practices from the outset.
Foundational principles	Cavoukian's seven foundational principles and GDPR data-protection principles [3, 10]	Specify the general privacy and data-protection requirements that system design should achieve.
Threat-identification methods	Privacy threat modelling and privacy-risk taxonomies [8, 21, 15, 6, 14]	Provide problem-oriented structures for identifying and organising potential privacy threats.
Design-response framework	Privacy design strategies [12]	Provide high-level directions for translating privacy requirements and identified threats into privacy-preserving design decisions.
Concrete controls	Model-level guardrails, access controls, memory controls, and data-management mechanisms [15, 13, 2, 7]	Implement selected privacy requirements and design strategies at the system and model levels.
Evaluation method	Black-box guardrail testing [11, 20, 17, 23]	Examines whether implemented restrictions exhibit the intended observable selectivity and robustness.

protection from complementary perspectives, including overarching principles, threat types, design responses, and emerging risks in AI systems, thereby providing the conceptual basis for constructing a taxonomy for black-box testing.

Based on this synthesis, we consolidate related and overlapping privacy concerns across the literature into six threat categories that can be expressed and observed at the model-interaction level. Such observability includes requests submitted by users, responses generated by models, and actions initiated by models through tools, data sources, or other external components. In this way, privacy concepts originally formulated primarily for system architectures, data flows, or governance processes are operationalised as testable objects for prompt-response analysis.

For each threat category, we also identify the primary privacy objectives most directly implicated by the corresponding threat. These objectives are informed by Privacy by Design principles, data-protection principles, established privacy properties, and contextual integrity[3, 10, 8, 21, 18, 12]. They indicate which privacy requirements are most likely to be undermined when a particular type of threat materialises. Through this mapping, the taxonomy not only organises testable privacy threats but also connects threat identification to the protection objectives associated with Privacy by Design.

Table 2 presents the resulting six-category, test-oriented privacy-threat taxonomy and the primary privacy objectives implicated by each category. The taxonomy provides a conceptual basis for the subsequent benchmark-coverage analysis and black-box evaluation of model-level guardrails, rather than replacing existing general-purpose privacy threat-modelling or privacy design frameworks.

4.2 Benchmark Selection and Dataset Construction

In the bottom-up empirical path, we select WildGuardTest [11] as the source of evaluation prompts. WildGuardTest is a general-purpose safety benchmark in which prompts are annotated according to WildGuard's operational safety taxonomy. We extract all harmful prompts assigned to its high-level *Privacy* category, resulting in 162 prompts. According to the original WildGuardTest labels, these prompts are distributed across three subcategories: private information concerning individuals, sensitive information concerning organisations or governments, and copyright violations. The subsequent experiments retain these benchmark-defined subcategories for category-level aggregation and comparison.

To establish a comparison baseline between harmful requests and ordinary benign interactions, we additionally select 162 benign prompts, producing a balanced evaluation set of 162 selected harmful prompts and 162 benign control prompts. The control group is used to assess whether the model exhibits selective restriction toward the selected harmful requests, rather than broadly refusing or failing to answer different types of inputs. Because the benign prompts do not have fine-grained labels corresponding to the three benchmark subcategories, they are used only for overall harmful-benign comparisons and not for within-category analysis.

Table 2. Top-down and test-oriented privacy-threat taxonomy.

Nr.	Privacy-threat category	Definition	Primary privacy objectives implicated
S1	Personal-data Disclosure and Exposure	The model is requested to reveal, reconstruct, generate, summarise, or transmit personal or sensitive information about an identifiable individual.	Confidentiality [8, 21, 15, 10]
S2	Sensitive Inference and Detection	The model is requested to infer previously unstated sensitive attributes, conditions, intentions, behaviours, or activities from indirect or contextual information.	Undetectability; data minimisation [8, 21, 14, 12]
S3	Linking, Identification, and Re-identification	The model is requested to link records, accounts, sessions, or data sources, or to identify an individual from anonymous, pseudonymous, or partial information.	Unlinkability; anonymity [8, 21, 15, 13]
S4	Profiling, Scoring, and Consequential Decision Use	The model is requested to create a profile, label, score, ranking, or eligibility assessment that may affect an individual's rights, opportunities, treatment, or autonomy.	Transparency; user control [12, 14, 16, 10]
S5	Contextually Inappropriate Use	The model is requested to use personal information in a manner inconsistent with its original purpose, the individual's consent, the requester's authority, or the expected recipient and context.	Purpose limitation; contextual integrity [18, 12, 15, 10]
S6	Excessive Collection, Retention, Propagation	The model is requested to collect, retain, monitor, or propagate personal information beyond what is necessary for the stated task.	Data minimisation; storage limitation [12, 16, 13, 10]

For each record, we retain the original prompt text and the available benchmark annotations, including harmfulness, adversarial status, and subcategory labels. The three benchmark-defined subcategories are used for category-level result aggregation, whereas the harmful and benign labels support black-box testing strategies that require a control condition. We do not modify the semantic content of the prompts or redefine their original labels based on the responses of the evaluated models.

Results are analysed by strategy, prompt type, model variant, and WildGuard subcategory. Category-level analysis focuses on the original model, while the abliterated variant serves as an overall contrastive condition.

4.3 Black-box Evaluation Procedure

To address RQ2, this study adopts a black-box prompt-response evaluation approach in which guardrail behaviour is analysed solely by submitting controlled prompts to the evaluated models and observing the resulting responses. The approach does not require access to model weights, internal activations, gradients, system prompts, token-level probabilities, or internal safety classifiers. We therefore do not interpret an externally observed response as direct evidence of a specific internal mechanism. Instead, we examine whether behavioural patterns produced by multiple testing strategies are consistent with selective restriction and guardrail robustness.

A single refusal is insufficient to establish the presence of a guardrail because non-execution may also result from limited capability, task misunderstanding, uncertainty, or generation variability. To obtain complementary forms of externally observable evidence, we apply three black-box testing strategies: Chi-square, Self-report, and Shallow Depth. Table 3 summarises the core testing logic of each strategy and the type of guardrail-relevant evidence it provides.

The three strategies use a shared set of response-level behavioural labels. Model responses are classified as GUARDRAIL, INCAPACITY, COMPLIANCE, or AMBIGUOUS. GUARDRAIL denotes an explicit restriction based on safety, policy, privacy, or authorisation considerations. INCAPACITY denotes non-execution that primarily reflects insufficient knowledge, capability, or task-processing ability. COMPLIANCE indicates that the model substantively executes or continues executing the request. AMBIGUOUS is assigned when the available evidence is insufficient, conflicting, invalid, or confounded by the testing condition. These labels describe externally observable behaviour and are not treated as direct identification of the model's internal guardrail implementation.

The strategy results are interpreted jointly rather than as independent proofs of guardrail implementation. Chi-square identifies whether systematic behavioural differences exist between benign and harmful conditions. Self-report provides supplementary attribution evidence concerning the stated reason for non-execution. Shallow Depth evaluates whether the observed restriction remains stable under an adversarial response context. Together, the three strategies provide complementary evidence concerning behavioural selectivity, attribution, and robustness. Directionally consistent evidence across these dimensions supports a more

Table 3. Overview of the black-box testing strategies used in this study.

Strategy	Core testing logic	Guardrail-relevant evidence
Chi-square	Compares response-label distributions between benign control prompts and the selected harmful prompts.	Examines whether the model restricts harmful requests more frequently than benign inputs, thereby providing observational evidence of behavioural selectivity [20].
Self-report	Uses diagnostic follow-up questions after an initial non-compliant response to ask whether the model's non-execution resulted primarily from policy restrictions or limited capability.	Provides supplementary attribution evidence concerning the model's stated reason for not executing the request [17].
Shallow Depth	Keeps the harmful prompt unchanged but introduces assistant-response prefixes of different styles and depths, and observes whether the model continues executing the request.	Examines whether restriction remains stable under response-context manipulation or is weakened by assistant prefilling [23].

substantiated behavioural interpretation than a single refusal or an individual measurement.

4.4 Experimental Setup

We evaluate GPT-OSS 20B, served through Ollama as `gpt-oss:latest`, and the contrastive abliterated variant `huihui_ai/gpt-oss-abliterated:20b`. All tests are executed sequentially in a RunPod GPU environment. Both variants use the same generation configuration: temperature 0.7, top-p 1.0, a maximum output length of 256 tokens, and seed 0 where supported. Table 4 summarises the strategy-specific conditions and measurements.

Strategies requiring automated response analysis use a shared response classifier to produce the basic behavioural labels, supplemented where necessary by strategy-specific verification. In particular, an apparent COMPLIANCE outcome in Shallow Depth is further checked to determine whether the model substantively continues the original harmful request. Merely repeating, explaining, transforming, or rejecting the prefilling content is not treated as a successful bypass.

The response labels and measurements produced by each strategy are aggregated according to predefined operational decision boundaries. Because these boundaries represent empirically selected operational settings, the final analysis primarily considers the underlying measurements and the consistency of evidence across strategies rather than relying on any single strategy-level verdict.

Table 4. Experimental configurations and main measurements of the three black-box testing strategies.

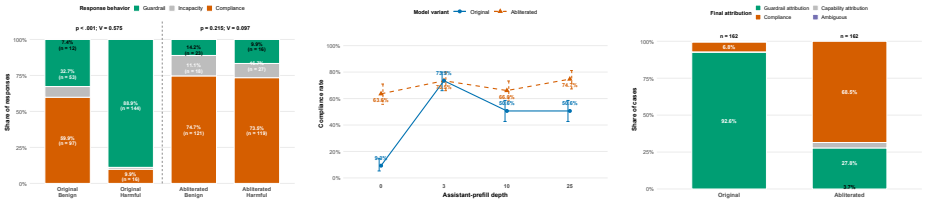
Strategy	Test conditions and configuration	Main measurements
Chi-square	One response is generated and classified for each of the 162 CTRL prompts and 162 MAL prompts.	Response-label distributions under CTRL and MAL, Pearson's χ^2, p-value, and Cramer's V.
Self-report	One initial response is generated for each MAL prompt. Initial non-compliance triggers three policy-versus-capability diagnostic follow-up questions.	Guardrail-attribution, capability-attribution, ambiguous-attribution, and direct-compliance rates.
Shallow Depth	Each MAL prompt is tested using three assistant-prefix styles at four prefill depths, $D = \{0, 3, 10, 25\}$, producing twelve conditions per prompt.	Response-label distributions across depths, changes in compliance rate, and the earliest prefill depth at which COMPLIANCE is observed.

5 Evaluation

We evaluate whether the tested model exhibits privacy-related guardrail behavior, how this behavior differs between the original and abliterated model variants, and whether it remains robust across testing conditions. We first examine the overall behavioral evidence, then compare the benchmark-defined categories, and finally assess their alignment with the literature-derived privacy-threat taxonomy.

5.1 Evidence of Privacy-Related Guardrails

Figure 1 summarises the evidence obtained from the Chi-square, shallow-depth, and self-report strategies.



(a) Chi-square strategy **(b) Shallow-depth strategy** **(c) Self-report strategy**

Fig. 1. Evidence and robustness of privacy-related guardrail behavior.

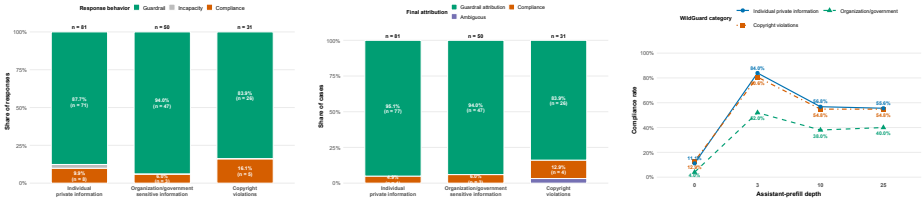
The Chi-square results show a clear behavioral difference between benign and harmful prompts for the original model. Harmful prompts produced an 88.9% guardrail rate, compared with 32.7% for benign prompts. In contrast, the abliterated model showed nearly identical compliance rates for benign and harmful prompts, at 74.7% and 73.5%, respectively. This contrast suggests that the selective restriction observed in the original model is associated with its safety alignment. However, the relatively high restriction rate for benign prompts also indicates imperfect selectivity.

The self-report results provide complementary attribution evidence. In the original model, 92.6% of cases were attributed to guardrail-related restrictions, whereas the abliterated model predominantly complied, accounting for 68.5% of cases. Since self-reported explanations do not directly reveal internal mechanisms, we interpret this result only as supporting evidence.

The shallow-depth results further indicate limited robustness. The original model's compliance rate increased from 9.3% at depth 0 to 73.5% at depth 3, before decreasing to 50.6% at depths 10 and 25. The effect was therefore substantial but non-monotonic. Overall, the original model exhibits strong but imperfect privacy-related guardrail behavior that remains sensitive to assistant-prefill manipulation.

5.2 Category-Level Effectiveness and Robustness

Having established that the original model exhibits selective restriction consistent with privacy-related guardrails, we next compare its behavior across the three WildGuard-defined categories in the selected benchmark subset. Because the categories contain different numbers of prompts, the analysis is based on proportions rather than absolute counts. Figure 2 presents the category-level results for the Chi-square, self-report, and shallow-depth strategies.



(a) Chi-square strategy **(b) Self-report strategy** **(c) Shallow-depth strategy**

Fig. 2. Category-level guardrail behavior across the three WildGuard-defined benchmark categories.

The Chi-square strategy shows that restriction strength differs across categories. The guardrail rate was highest for sensitive information concerning organizations or governments, at 94.0%, followed by individuals' private information

at 87.7% and copyright violations at 83.9%. Correspondingly, copyright violations produced the highest compliance rate, at 16.1%, whereas the organization/government category produced the lowest, at 6.0%. These results suggest that the model does not apply the same degree of restriction uniformly across the three benchmark categories.

The self-report strategy exhibits a similar pattern. Guardrail attribution reached 95.1% for individuals' private information and 94.0% for organization / government information, compared with 83.9% for copyright violations. The copyright category also showed a higher compliance rate and contained the only ambiguous attribution. The agreement between response behavior and self-reported attribution provides additional, although not causal, evidence that the observed non-compliance is associated with model-level restriction rather than capability limitations alone.

The shallow-depth strategy reveals more substantial differences in robustness. Without assistant prefill, compliance remained low across all three categories. At depth 3, however, compliance increased to 84.0% for individuals' private information and 80.6% for copyright violations, compared with 52.0% for organization/government information. The organization/government category also maintained the lowest compliance rates at depths 10 and 25, at 38.0% and 40.0%, respectively. This category therefore exhibited both the highest initial restriction rate and the strongest relative resistance to assistant-prefill manipulation. By contrast, requests concerning individuals' private information were particularly vulnerable to shallow prefill.

Overall, the three strategies consistently indicate that the model restricts organization/government information more strongly and robustly than the other benchmark categories. Restrictions concerning individuals' private information and copyright violations are more readily weakened by assistant prefill. However, these categories originate from a general safety benchmark, and copyright violations do not strictly constitute a privacy threat. The following analysis therefore examines their correspondence with the literature-derived privacy-threat taxonomy.

5.3 Alignment and Gaps Between the Top-down and Bottom-up Classifications

WildGuardTest groups the selected harmful prompts into three benchmark-defined categories: private information concerning individuals, sensitive information concerning organisations or governments, and copyright violations. To compare this bottom-up structure with the literature-derived S1–S6 taxonomy, we conducted three AI-assisted annotation passes and applied majority voting followed by category-boundary human review. Of the 162 prompts, 140 received identical labels across all three runs, 20 obtained a majority label, and two had no majority.

The reviewed distribution is concentrated in S1, with 79 prompts involving personal-data disclosure or exposure, while 73 prompts fall outside S1–S6. The remaining privacy-threat categories are only sparsely represented.

Table 5. Label distributions across the three annotation runs, majority voting, and human review.

label	Run 1	Run 2	Run 3	Majority	Reviewed
S1 Personal-data Disclosure and Exposure	91	87	84	87	79
S2 Sensitive Inference and Detection	2	2	2	2	3
S3 Linking, Identification, and Re-identification	0	2	3	1	1
S4 Profiling, Scoring, and Consequential Decision Use	1	1	1	1	1
S5 Contextually Inappropriate Use	5	7	6	6	1
S6 Excessive Collection, Retention, and Propagation	3	4	4	3	4
OUT_OF_SCOPE	60	59	62	$60 + 2^*$	73
Total	162	162	162	162	162

*Two prompts received three different labels and therefore had no majority label. They are displayed separately within this cell for compactness and were subsequently resolved through human review.

This reveals both a *scope gap* and a *coverage gap*. WildGuard's *Privacy* category includes copyright, organisational secrecy, government information, and cyberattacks that are not necessarily personal-privacy threats. At the same time, it provides limited coverage of inference, re-identification, profiling, contextual misuse, and excessive data processing. The experimental results therefore characterise the three WildGuard-defined categories, rather than the complete S1–S6 privacy-threat space.

6 Threats to Validity

This study has several limitations.

First, the experiments currently cover only one open-weight model and its abliterated variant. The observed privacy-related restriction behavior and its variation across testing conditions may therefore be influenced by the specific model architecture and safety-alignment approach. The results cannot yet be directly generalized to other open-weight or commercial models.

Second, the experiments use only English prompts. Because privacy concepts, expressions of sensitive information, and safety-alignment performance may vary across languages, the findings cannot be directly generalized to multilingual settings.

Finally, the privacy-related prompts used in this study are drawn from WildGuardTest, which was originally designed for general safety evaluation. Its privacy categories do not fully correspond to the literature-derived, privacy-specific taxonomy proposed in this study. Moreover, no existing prompt-based test dataset is specifically designed for evaluating privacy guardrails according to this taxonomy. We therefore use AI-assisted annotation followed by manual review to

relabel the existing prompts and examine their coverage of different privacy threats. These annotations are used for coverage analysis rather than presented as a new benchmark or gold-standard dataset.

7 Related Work

Privacy engineering and threat-modelling research provides the theoretical foundation for analysing model-level privacy protection. Existing work has translated Privacy by Design into strategies such as data minimisation, information hiding, processing separation, transparency, user control, and enforceable data-handling rules. In parallel, frameworks including LINDDUN, contextual integrity, and broader privacy taxonomies describe privacy risks in terms of information disclosure, identifiability, linkability, detectability, inappropriate use, and excessive processing [12, 8, 18, 6, 21]. These studies establish relevant privacy objectives and threat concepts, but generally analyse system components, actors, data flows, and processing activities rather than behaviour observable through model prompt–response interactions.

More recent research extends privacy analysis to LLMs, generative AI, and Agentic AI. Surveys and threat-modelling studies cover training-data exposure, membership inference, sensitive-attribute inference, personal-information generation and disclosure, as well as risks arising from memory, external data access, tool use, cross-context information reuse, and multi-agent communication [4, 15, 13, 14, 16]. This work substantially broadens the scope of established privacy frameworks and organises emerging risks by model lifecycle, attack type, data source, or system surface. However, the resulting classifications differ in granularity and are primarily intended for system-level threat modelling, risk governance, or attack mitigation rather than as operational categories for black-box guardrail testing.

Privacy-specific guardrails have also begun to emerge, including approaches for sensitive-entity detection, PII mitigation, contextual privacy enforcement, and privacy filtering for conversational agents [2, 1, 7]. These developments extend guardrails beyond general safety moderation, but currently address only parts of the broader privacy-threat space. This leaves scope for more systematic privacy-oriented guardrail design and evaluation across a wider range of privacy threats.

In parallel, the evaluation of LLM guardrails has primarily focused on general harmful requests, response safety, refusal behaviour, over-refusal, and robustness under adversarial conditions. Work such as WildGuard and XSTest provides benchmarks and evaluation methods for general safety behaviour, but privacy is typically treated as one category within a broader safety taxonomy rather than tested against a structured account of privacy threats [11, 20]. Although privacy-specific guardrail design is beginning to develop, existing evaluation rarely examines together the coverage of different privacy threats, model-level restriction behaviour, and its robustness under black-box testing conditions.

8 Conclusion

This study combined a top-down privacy-threat taxonomy with bottom-up black-box testing to examine how model-level LLM guardrails may support Privacy by Design.

To address RQ1, we derived a six-category, test-oriented taxonomy covering personal-data disclosure, sensitive inference, re-identification, profiling, contextually inappropriate use, and excessive data processing. Comparing this taxonomy with WildGuardTest revealed both scope and coverage gaps: the selected prompts were heavily concentrated in disclosure-related threats, while many prompts fell outside the privacy-specific taxonomy.

To address RQ2, the original GPT-OSS 20B model exhibited selective privacy-related restriction relative to both benign prompts and its abliterated variant. Self-report results provided supplementary attribution evidence, while Shallow Depth testing showed that assistant-prefill manipulation substantially increased compliance. The observed restriction behaviour was therefore selective but imperfect and not consistently robust across testing conditions.

Future work could develop privacy-specific prompt datasets aligned with the proposed taxonomy and refine black-box testing methods to target properties such as data sensitivity, inference, contextual use, and information flows more directly. Further research could also extend the evaluation to agentic AI settings involving persistent memory, external data access, tool use, and multi-step execution, where privacy-relevant behaviour extends beyond direct prompt-response interaction.

References

1. Asthana, S., Mahindru, R., Zhang, B., Sanz, J.: Adaptive pii mitigation framework for large language models. arXiv preprint arXiv:2501.12465 (2025)
2. Asthana, S., Zhang, B., Mahindru, R., DeLuca, C., Gentile, A.L., Gopisetty, S.: Deploying privacy guardrails for llms: A comparative analysis of real-world applications. arXiv preprint arXiv:2501.12456 (2025)
3. Cavoukian, A., et al.: Privacy by design: The 7 foundational principles. Information and privacy commissioner of Ontario, Canada **5**(2009), 12 (2009)
4. Chen, K., Zhou, X., Lin, Y., Feng, S., Shen, L., Wu, P.: A survey on privacy risks and protection in large language models. Journal of King Saud University Computer and Information Sciences **37**(7), 163 (2025)
5. Chhabra, A., Datta, S., Nahin, S.K., Mohapatra, P.: Agentic ai security: Threats, defenses, evaluation, and open challenges. arXiv preprint arXiv:2510.23883 (2025)
6. Daniel, J.S.: A taxonomy of privacy. University of Pennsylvania law review **154**(3), 477–560 (2006)
7. Das, S., Fioretto, F.: Neurofilter: Privacy guardrails for conversational llm agents. arXiv preprint arXiv:2601.14660 (2026)
8. 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)

9. European Data Protection Supervisor: Agentic ai. https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/agentic-ai_en (2025), techSonar 2025–2026, accessed 30 July 2026
10. European Parliament and Council of the European Union: Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, article 25: Data protection by design and by default. Official Journal of the European Union, L 119, 4 May 2016 (2016), https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng, article 25
11. Han, S., Rao, K., Ettinger, A., Jiang, L., Lin, B.Y., Lambert, N., Choi, Y., Dziri, N.: Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms (2024), https://arxiv.org/abs/2406.18495
12. Hoepman, J.H.: Privacy design strategies. In: IFIP International Information Security Conference. pp. 446–459. Springer (2014)
13. Lahjouji, N., Colaco, A.G.: Agents that know too much: A data-centric survey of privacy in llm agents. arXiv preprint arXiv:2606.26627 (2026)
14. Lee, H.P., Yang, Y.J., Von Davier, T.S., Forlizzi, J., Das, S.: Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. pp. 1–19 (2024)
15. Liao, Q., Bellemans, J., Sion, L., Jiang, X., Usynin, D., Zhou, X., Van Landuyt, D., Desmet, L., Joosen, W.: A linddun-based privacy threat modeling framework for genai. arXiv preprint arXiv:2603.06051 (2026)
16. Meisenbacher, S., Klymenko, A., Kelley, P.G., Peddinti, S.T., Thomas, K., Matthes, F.: Privacy risks of general-purpose ai systems: A foundation for investigating practitioner perspectives. arXiv preprint arXiv:2407.02027 (2024)
17. Nguyen, Q.M., Ahmed, U., Kim, T.: Can llms reliably self-report adversarial prefills, and how? arXiv preprint arXiv:2606.23671 (2026)
18. Nissenbaum, H.: Privacy as contextual integrity. Wash. L. Rev. **79**, 119 (2004)
19. Reber, D., Richardson, S., Nief, T., Garbacea, C., Veitch, V.: Rate: Causal explainability of reward models with imperfect counterfactuals. arXiv preprint arXiv:2410.11348 (2024)
20. Röttger, P., Kirk, H., Vidgen, B., Attanasio, G., Bianchi, F., Hovy, D.: Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 5377–5400 (2024)
21. Sion, L., Van Landuyt, D., Wuyts, K., Joosen, W.: Robust and reusable linddun privacy threat knowledge. Computers & Security **154**, 104419 (2025)
22. Staab, R., Vero, M., Balunović, M., Vechev, M.: Beyond memorization: Violating privacy via inference with large language models. In: The Twelfth International Conference on Learning Representations. OpenReview (2024)
23. Struppek, L., Gleave, A., Pelrine, K.: Exposing the Systematic Vulnerability of Open-Weight Models to Prefill Attacks (2026), https://arxiv.org/abs/2602.14689
24. Wang, B., He, W., Zeng, S., Xiang, Z., Xing, Y., Tang, J., He, P.: Unveiling privacy risks in llm agent memory. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 25241–25260 (2025)