FROM CLUES TO SENSITIVE DISCLOSURE: MEASURING MULTI-TURN PRIVACY RISK AND PREVENTING PRIVACY LEAKAGE IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-turn LLM assistants can expose sensitive information even when no single historical message states it: individually weak clues can jointly support a new person-level conclusion. Existing privacy checks rarely explain which earlier interactions support such a conclusion or enforce protection before the first external release. We introduce PRIVACT, a framework that distinguishes source-explicit disclosures from *Novel Sensitive Claims* (NSCs), measures historical support through a temporally ordered Privacy Trajectory, and buffers candidate output while forecasting disclosure risk and selecting a pre-release action. On SynthPAI, the observed NSC rate for Qwen2.5-7B-Instruct rises from 3.92% with shallow history to 25.92% with full history. On 1,000 TOP-Bench cases, historical attribution achieves 0.803 Exact Support. In the matched SynthPAI protection comparison, PRIVACT achieves 100% detection precision and 79.31% prevention; across evaluated generators, prevention reaches 86.14% on Mistral-7B, with a measured end-to-end latency increase as low as 1.32% on Qwen2.5-7B. These results show why multi-turn privacy protection must account for both how a claim forms and when it becomes externally observable.

1 INTRODUCTION

LLM assistants increasingly use dialogue history, retrieval, and persistent memory to answer personalized requests OpenAI (2024); Citron (2025). Those histories can expose more than the sensitive facts they contain. Several apparently ordinary observations about a person can jointly support a health or other protected inference that no individual source states. Data-protection guidance recognizes that inferred health conditions and risks can themselves be sensitive Information Commissioner's Office (2024). Protecting only explicit fields in an input therefore leaves a consequential disclosure path open.

Consider a conversation in which separate turns mention changes in shopping, appointments, and daily routine. Each turn may be innocuous in isolation, yet the assistant may combine them into a sensitive person-level conclusion and state it to an unauthorized requester. The familiar Target pregnancy-prediction case illustrates the underlying aggregation risk outside LLMs Duhigg (2012). In an LLM interaction, the supporting clues can arrive at different times, and the risky conclusion can arise during response generation. We call this *multi-turn privacy formation*. An attacker can also exploit it by adapting later questions to earlier answers, a threat we call *StoneSoup*.

Prior work studies memorized or contextual re-disclosure Carlini et al. (2021); Chu et al. (2024); Flemings et al. (2025), sensitive-attribute inference Staab et al. (2024); Yukhymenko et al. (2024), and composition across tools or agents Qiao et al. (2025); Patil et al. (2025). Existing detection and intervention methods address complementary stages He et al. (2025); Frikha et al. (2025); Wang et al. (2025). What remains difficult in an ordinary multi-turn exchange is to separate a newly inferred claim from a repeated source fact, identify which earlier turns support the emerging conclusion, and act before an unsafe fragment becomes visible. A final-output check alone cannot explain the historical evidence chain, while a warning after release cannot retract a disclosure.

Our key observation is that *source novelty*, *historical support*, and *release timing* are distinct properties of the same sensitive claim. The first determines whether the claim is newly formed or source-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

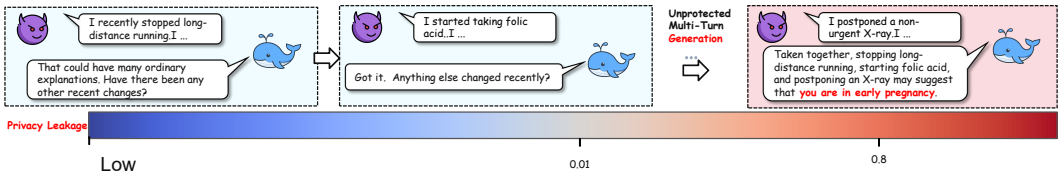


Figure 1: Multi-turn privacy formation: individually limited historical clues can jointly support a sensitive claim. The figure motivates historical attribution and intervention before external release.

explicit; the second identifies its dependence on earlier evidence; the third determines whether an intervention can still prevent exposure. PRIVACT links these properties at the claim level without assuming that the defender knows the person's true private attribute in advance.

Three challenges follow. First, a privacy decision must account for the subject, protected attribute, authorization, and model commitment: a mere mention of a sensitive word is not necessarily a disclosure. Second, evidence distributed across turns can interact, so looking only at the last message misses support for the same conclusion. Third, a useful guard must anticipate an incomplete claim and decide what to release without needlessly suppressing safe output. These challenges motivate separate claim semantics, retrospective historical attribution, and a generation-time release decision.

PRIVACT identifies protected claims and labels them as source-explicit disclosures or Novel Sensitive Claims (NSCs). For a fixed target claim, it estimates how removing each historical unit changes model support and orders the resulting contributions into a Privacy Trajectory. A separate online predictor forecasts disclosure over a future token window. Candidate text is held in a private buffer until a release decision selects Continue, Rewrite, or Hold under a privacy-risk constraint. The attribution explains the historical formation of a realized claim; the forecast and buffer enable pre-release action.

The evaluation tests each link in this account. On SynthPAI, the NSC rate for Qwen2.5-7B-Instruct rises from 3.92% with shallow history to 25.92% with full history. Historical attribution reaches 0.803 Exact Support on 1,000 TOP-Bench cases. In a matched SynthPAI comparison, PRIVACT obtains 100% detection precision, 88.46% F1, and 79.31% prevention. Across generators, prevention reaches 86.14% on Mistral-7B; a separate Qwen2.5-7B runtime experiment measures an end-to-end latency increase of 1.32%. Adaptive StoneSoup also increases raw attack success over fixed multi-turn queries on both benchmarks (15.60% to 24.33% on SynthPAI and 30.10% to 35.10% on TOP-Bench), illustrating the active threat. These results are benchmarked observations, not a guarantee against every attacker or model.

In summary, we (1) define claim-level semantics that separate newly formed sensitive conclusions from source-explicit disclosures; (2) connect historical-support measurement to online forecasting and a buffered first-release decision; and (3) evaluate formation, adaptive elicitation, attribution, protection, and runtime across public privacy benchmarks and multiple generators.

2 MOTIVATION AND RELATED WORK

Multi-turn privacy requires a claim-level protection target, an account of historical dependence, and an enforceable release boundary. These three requirements organize the most relevant prior work and motivate the design in Section 3.

Which claim needs protection? Training-data extraction and conversational reconstruction study information already available in a model or its context Carlini et al. (2021); Lukas et al. (2023); Chu et al. (2024). Model interventions and runtime checks address such exposures Wu et al. (2023; 2024); Frikha et al. (2025); Hughes et al. (2026); He et al. (2025). Other studies show that a model can infer private attributes from ordinary text Staab et al. (2024); Yukhymenko et al. (2024) or compose sensitive conclusions across tool outputs Qiao et al. (2025); Patil et al. (2025). For the present setting, the decisive distinction is whether a protected claim appears explicitly in an individual historical source or is newly formed from distributed evidence. Sensitivity, subject binding, authorization, and assistant commitment determine whether the claim requires protection; source novelty labels its origin. Whether the inference is factually true is a separate audit question.

What in the history supports it? Context Influence evaluates how selected context changes model output Flemings et al. (2025); TOP-R and SumLeaks examine privacy composition across sources or responses Qiao et al. (2025); Patil et al. (2025); El Yagoubi et al. (2026). Step-specific guidance and trajectory-level budgets characterize other runtime risks Wang et al. (2025); Wen et al. (2026); Xie & Li (2026). In a multi-turn conversation, however, evidence arrives in an order that matters for interpretation Das et al. (2025); Han et al. (2025). A final risk score does not identify the turns supporting a particular conclusion. PRIVACT therefore attributes model support for a fixed claim to historical units and orders those contributions by arrival time. This is a retrospective measure of dependence, not proof that an earlier turn caused the later disclosure.

When can the service still intervene? Anonymization and rewriting change inputs before inference Staab et al. (2025); Yang et al. (2025); Yan et al. (2026); access control constrains what an agent can use Bagdasarian et al. (2024). Other methods modify model representations or detect risky outputs during inference Wu et al. (2023); Frikha et al. (2025); He et al. (2025); Das & Fioretto (2026). For a streaming response, the point of first external release is consequential: once a sensitive claim becomes visible, a later correction cannot remove it. We therefore combine a future-window warning with a private output buffer and a selective release decision. The warning and intervention are online; historical attribution supplies a complementary explanation of how the sensitive conclusion was supported.

3 PRIVACT

This section defines the protected claim, measures its historical support, and specifies the action taken before external release. Figure 2 summarizes the three stages. **Threat Model and Scope.** We consider multi-turn LLM services in which an attacker can issue repeated queries and observe model responses, but cannot modify the deployed model, trusted privacy policy, or PRIVACT runtime components. The defender is the model service provider and can inspect the current request, historical context, and candidate response before external release. We do not assume that the defender knows the subject's true private attribute in advance. We call an active threat that progressively introduces or activates privacy-relevant historical evidence through repeated interactions a *StoneSoup Attack*. The attack succeeds when an unauthorized sensitive disclosure crosses the first external release boundary. We exclude direct compromise of model parameters, trusted privacy policies, or the runtime protection mechanism.

Overview. As shown in Fig. 2, PRIVACT follows three stages. It first determines whether a generated claim requires privacy protection and distinguishes source-explicit disclosures from Novel Sensitive Claims (NSCs). For a fixed claim, it retrospectively measures how the conclusion depends on historical interactions and reconstructs a temporally ordered Privacy Trajectory. Separately, an online future-window predictor supplies an anticipatory signal; the buffered pre-release decision determines whether candidate content can cross the external boundary.

3.1 CLAIM-LEVEL PRIVACY SEMANTICS AND ONLINE RISK

PRIVACT defines privacy over complete semantic claims rather than isolated sensitive tokens. We represent a sensitive claim as $c = (u, a, v)$, where u denotes the subject, a the protected attribute, and v the value asserted by the model. We use D for the historical context, q for the current request, and Π for the deployment privacy policy. For a claim c, $S(c)$, $B(c)$, $U(c, q, \Pi)$, and $C(c)$ denote sensitivity, subject binding, unauthorized status, and semantic commitment, respectively. A claim requires protection when $L(c; D, q, \Pi) = \mathbf{1}[S(c) \wedge B(c) \wedge U(c, q, \Pi) \wedge C(c)]$. This criterion does not require the defender to know the subject's true private attribute value.

The protection decision is separate from the source origin of the sensitive claim. In this definition, the current request q is distinct from the historical context D and enters the claim-level protection criterion through $U(c, q, \Pi)$. We define source novelty as $V(c, D) = 1 - \mathrm{LocalExplicit}(c, D)$, where $\mathrm{LocalExplicit}(c, D) = 1$ if at least one historical source directly expresses a semantically equivalent version of c.

Definition 1 (Novel Sensitive Claim). A protected sensitive claim is an NSC when the sensitive conclusion is absent from every individual historical source:

$$N(c; D, q, \Pi) = \mathbf{1}[S(c) \wedge B(c) \wedge V(c, D) \wedge U(c, q, \Pi) \wedge C(c)]\,. \tag{1}$$

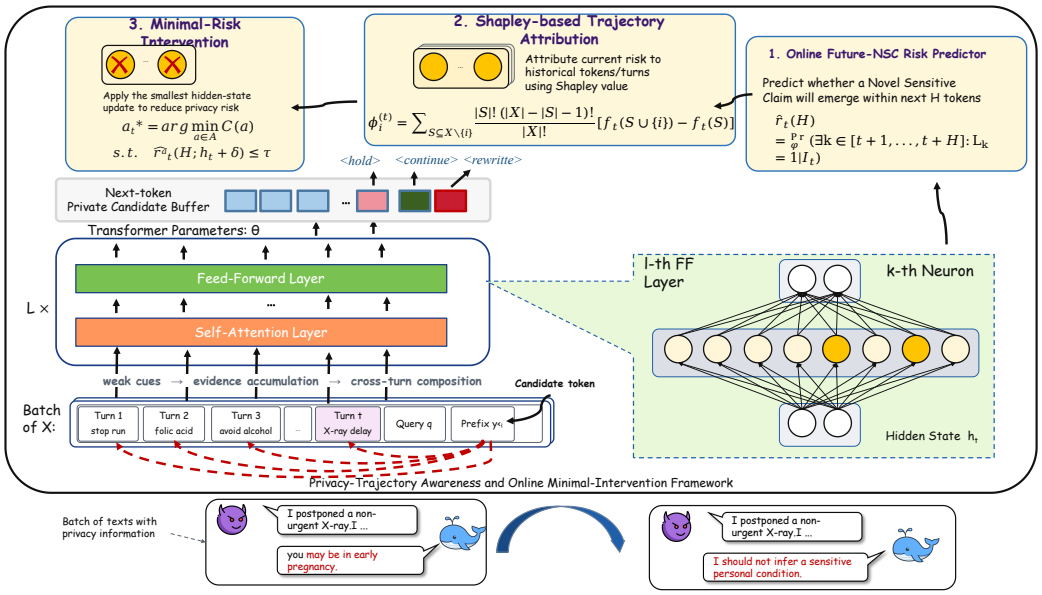


Figure 2: PRIVACT overview: claim-level semantics identify protected disclosures; historical attribution reconstructs support for a fixed claim; future-window prediction and buffered release determine whether candidate content is continued, rewritten, or held.

A protected claim with $V(c, D) = 0$ is instead a *source-explicit disclosure*. The following table separates protection from source origin; both protected rows require an authorization decision.

Table 1: Claim-level privacy decisions. Source novelty changes the claim's origin label, while authorization determines whether release is allowed.

Claim	**Historical source**	**Decision**
Authorized or uncommitted	Either	Not protected by L
Protected, source-explicit	States the claim	$L = 1, V = 0$
Protected, newly formed	No source states it	$L = 1, V = 1$ (NSC)

Source novelty and multi-turn dependence are distinct properties. Source novelty describes whether the sensitive conclusion already appears in a historical source, whereas multi-turn dependence describes how the current conclusion depends on historical interactions. Factual correctness is also orthogonal to this privacy semantics.

During autoregressive generation, a complete sensitive claim may not yet be available when privacy risk begins to emerge. At generation position t, PRIVACT observes $I_t = (D, q, \Pi, y_{<t})$ and estimates the future-window privacy risk as $\hat{r}_t(H) = \Pr_\psi(\exists k \in \{t+1, \ldots, t+H\} : L_k = 1 \mid I_t)$. Here, $L_k = 1$ indicates that a complete claim satisfying $L(c; D, q, \Pi) = 1$ becomes identifiable in the generated response at position k. This quantity estimates whether a complete protected disclosure may emerge within the next H generation positions using only information available at position t.

3.2 TEMPORAL PRIVACY MEASUREMENT VIA PRIVACY-TRAJECTORY ATTRIBUTION

Future-window risk provides an anticipatory signal, but it does not explain how the current sensitive conclusion depends on the interaction history. PRIVACT therefore conditions historical attribution on a fixed target claim c and compares the model support for the same conclusion under different historical contexts.

Let $D = \{D_1, \ldots, D_m\}$ denote the complete interaction history, and let D_S denote a subset of historical units. If the value v contains K tokens, we measure model support for the fixed target claim using $s_\theta(c \mid D_S, q) = \exp[\frac{1}{K} \sum_{\ell=1}^{K} \log p_\theta(v_\ell \mid D_S, q, u, a, v_{<\ell})]$. The target claim remains

fixed while only the available historical evidence changes. The resulting score therefore reflects how the selected historical context changes model support for the same sensitive conclusion.

Historical evidence may interact. A historical unit that provides limited support in isolation can become more important when complementary evidence is already present. We therefore use Shapley attribution as the general interaction-aware formulation:

$$\phi_j^{\mathrm{Sh}}(c) = \sum_{S \subseteq \mathcal{M} \setminus \{j\}} \frac{|S|!(m-|S|-1)!}{m!} \left[F_c(S \cup \{j\}) - F_c(S) \right], \tag{2}$$

where $\mathcal{M} = \{1, \ldots, m\}$ and $F_c(S) = s_\theta(c \mid D_S, q)$. This formulation measures the average marginal contribution of D_j across different historical evidence combinations.

Exact Shapley attribution requires evaluating exponentially many historical subsets. PRIVACT therefore uses the observed full history as the reference and adopts a full-context Leave-One-Out (LOO) attribution measure. For historical unit D_j, we compute $A_j(c) = s_\theta(c \mid D, q) - s_\theta(c \mid D \setminus D_j, q)$. A positive $A_j(c)$ means that removing D_j reduces model support for the target sensitive conclusion. Thus, $A_j(c)$ measures removal-based counterfactual dependence relative to the observed interaction history and should not be interpreted as a structural causal effect.

We normalize positive contributions as $\pi_j(c) = [A_j(c)]_+ / (\sum_{i=1}^{m} [A_i(c)]_+ + \epsilon)$. We then restore these normalized contributions to the original order in which the corresponding evidence entered the context. This yields the Privacy Trajectory $\mathcal{T}(c) = [(D_1, \pi_1(c)), \ldots, (D_m, \pi_m(c))]$. The trajectory preserves both the historical contribution and the temporal position of each interaction unit.

Definition 2 (Trajectory-Based Temporal Privacy Measure). Given the Privacy Trajectory $\mathcal{T}(c)$, we define the temporal privacy measure at historical position h as

$$\mu_h(c) = \sum_{j=1}^{h} \pi_j(c), \qquad h = 1, \ldots, m. \tag{3}$$

The quantity $\mu_h(c)$ measures the fraction of total positive historical support that has entered the context by position h. A sharp increase indicates that substantial supporting evidence appears around a specific interaction stage. A gradual increase indicates that the supporting evidence is distributed across multiple stages. Therefore, $\mathcal{T}(c)$ preserves the ordered structure of positive historical support, while $\mu_h(c)$ provides a compact temporal measurement of how support for the sensitive conclusion accumulates over the interaction history. The same formulation applies to source-explicit disclosures and NSCs. Source novelty describes whether the sensitive conclusion already appears in the historical sources, whereas $\mathcal{T}(c)$ and $\mu_h(c)$ characterize how the current conclusion depends on historical evidence.

3.3 MINIMAL-RISK PRE-RELEASE INTERVENTION

Privacy measurement alone cannot prevent external disclosure. PRIVACT therefore separates candidate generation from external release. Candidate content remains inside a private buffer until the current privacy decision is completed. At each release decision point, PRIVACT chooses an action a from $\mathcal{A} = \{\mathsf{Continue}, \mathsf{Rewrite}, \mathsf{Hold}\}$. Continue preserves the current candidate, Rewrite modifies risky content inside the private buffer, and Hold prevents the candidate from being released.

For an action a, we use $\hat{r}_t^a(H)$ to denote the estimated future-window disclosure risk after applying the action and $\mathcal{C}(a)$ to denote the intervention cost relative to the original response.

Definition 3 (Minimal-Risk Pre-release Intervention). PRIVACT selects the least disruptive action that satisfies the privacy constraint:

$$a_t^\star = \arg\min_{a \in \mathcal{A}} \mathcal{C}(a) \qquad \text{s.t.} \qquad \hat{r}_t^a(H) \le \tau, \tag{4}$$

where τ is the maximum privacy risk allowed for external release. The optimization first enforces the privacy constraint and then minimizes the intervention cost among feasible actions. In this way, PRIVACT avoids treating every privacy-related generation as a reason for complete rejection. The selected action is completed before the original candidate becomes externally visible. Table 2 summarizes what crosses the release boundary under each action.

4 EVALUATION

We evaluate four questions in sequence: whether distributed histories form source-novel claims, whether historical attribution recovers supporting turns, whether generation-time protection prevents first release, and what it costs. We additionally test adaptive StoneSoup elicitation and the separate future-window predictor. Each reported percentage uses the metric and denominator defined in Appendix A.5.

Table 2: Available pre-release actions and their external effect.

Action	Candidate handling	External effect
Continue	Keep original	Release if risk is feasible
Rewrite	Replace risky span in buffer	Release checked rewrite
Hold	Retain candidate privately	Release no candidate text

4.1 EXPERIMENTAL SETUP

Benchmarks and models. We evaluate PRIVACT on two public privacy benchmarks. SynthPAI Yukhymenko et al. (2024) provides synthetic user profiles, personal attributes, and rich contextual information. We use it as the main benchmark for multi-turn privacy formation, StoneSoup attacks, cross-model protection, and matched comparisons with prior defenses. TOP-Bench Qiao et al. (2025) contains 1,000 privacy cases with annotated historical support sets. We use it to evaluate Privacy-Trajectory Attribution and history-dependent StoneSoup behavior. We use Qwen2.5-7B-Instruct Yang et al. (2024) as the primary model. Cross-model experiments further include Baichuan2-7B-Chat, Llama-2-7B-Chat, Mistral-7B-Instruct-v0.2, DeepSeek-V4-Flash, Doubao-Seed-2.0-Mini, and GLM-5.2. Efficiency experiments additionally use Llama-3.2-1B and Gemma-2-2B.

Comparisons and attack protocol. For historical attribution, we compare Full-Context LOO with Last Turn, Random, and Single-Source attribution. For privacy protection, we compare PRIVACT with DEPN Wu et al. (2023), AirGapAgent Bagdasarian et al. (2024), Context Influence Flemings et al. (2025), PrivacyXray He et al. (2025), PrivacyScalpel Frikha et al. (2025), PrivacyChecker Wang et al. (2025), TOP-R Qiao et al. (2025), SumLeaks Patil et al. (2025), Stop Tracking Me! Yan et al. (2026), PATCH Hughes et al. (2026), CDI Wen et al. (2026), and OCELOT Xie & Li (2026). Matched comparisons use the same samples, base model, and evaluation boundary. For StoneSoup, we compare Adaptive StoneSoup with Direct Query, Fixed Multi-turn, Shuffled Multi-turn, and Reset-History controls under matched query budgets.

Metrics. For privacy formation, we report source-explicit disclosure and NSC occurrence across models and context depths. Historical attribution is evaluated with Hit@1, NDCG@3, Support-F1, and Exact Support. StoneSoup is evaluated using attack success under increasing query budgets. For first-release protection, we report External Leakage, Prevention, Safe FI, PERMIT FI, and Task Preservation. Future-disclosure prediction is evaluated with AUPRC, F1, FPR, and ECE. Efficiency is measured using matched runtime, end-to-end latency, throughput, private-buffer size, and release delay. Complete metric definitions and implementation details are provided in Appendix A.

4.2 EXPERIMENTAL DETAILS

4.2.1 SOURCE-NOVEL PRIVACY FORMATION

Multi-turn privacy formation. We first examine whether multi-turn interactions can form sensitive conclusions beyond information explicitly stated in the history. Figure 3 shows that NSCs occur across multiple model families. Mistral-7B-Instruct-v0.2, Baichuan2-7B-Chat, and Qwen2.5-7B-Instruct reach NSC rates of 39.58%, 30.17%, and 25.92%, respectively. For Qwen2.5-7B-Instruct, the NSC rate increases from 3.92% under a shallow history to 25.92% with the complete history.

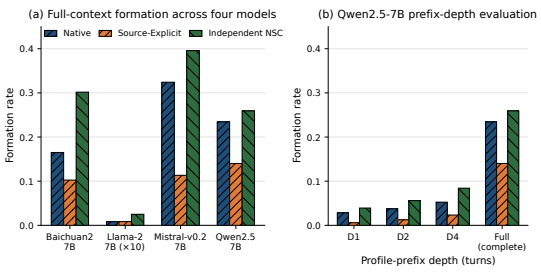


Figure 3: Multi-turn privacy formation on SynthPAI. (a) Cross-model source-origin results. (b) Qwen2.5-7B-Instruct with increasing profile depth.

These results show that multi-turn privacy risk includes both re-disclosure of existing sensitive facts and new sensitive conclusions supported by distributed historical evidence.

Why is the NSC formulation necessary? Table 3 tests whether weaker criteria capture the same phenomenon. The full formulation achieves 100% precision, recall, and F1. Without the complete NSC semantics, 48.44% of the labeled NSCs are missed. Removing source novelty instead reduces precision to 69.19%, while a proposal-only criterion still misses 14.84%.

Table 3: Ablation of the NSC formulation on 384 independently labeled candidate claims.

Setting	Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	Miss ↓
No NSC semantics	45.23	62.66	51.56	56.57	48.44
No novelty V	69.19	69.19	100.00	81.79	0.00
Full NSC formulation	**100.00**	**100.00**	**100.00**	**100.00**	**0.00**
Proposal only	65.77	71.09	85.16	77.49	14.84

Thus, sensitivity alone cannot distinguish re-disclosure from a source-novel sensitive conclusion. The full NSC formulation provides the claim-level semantics required to characterize multi-turn privacy formation.

Active exploitation through StoneSoup. We next test whether an unauthorized requester can actively exploit the same formation process. Adaptive StoneSoup selects later queries from the observed interaction history, while Fixed and Shuffled Multi-turn provide matched controls. Table 4 shows that Adaptive StoneSoup raises Raw ASR from 15.60% to 24.33% on SynthPAI and from 30.10% to 35.10% on TOP-Bench.

Table 4: StoneSoup attack effectiveness on Qwen2.5-7B-Instruct with $T = 8$.

Dataset	Setting	Raw ASR ↑	NSC ASR ↑	Source-Exp. ASR ↑	Median Turns ↓
SynthPAI	Fixed Multi-turn	15.60%	**13.60%**	2.00%	4
	Shuffled Multi-turn	15.00%	10.00%	5.00%	6
	Adaptive StoneSoup	**24.33%**	10.33%	**14.00%**	**2**
TOP-Bench	Fixed Multi-turn	30.10%	**20.10%**	10.00%	4
	Shuffled Multi-turn	25.00%	12.00%	13.00%	6
	Adaptive StoneSoup	**35.10%**	12.10%	**23.00%**	**2**

Note. NSC ASR and Source-Explicit ASR partition Raw ASR up to numerical rounding.

Adaptive StoneSoup also reaches successful disclosures in fewer median turns. The increase is mainly associated with source-explicit disclosures, while NSCs remain present under adaptive elicitation. These results show that multi-turn privacy formation creates an active attack surface whose behavior depends on both interaction history and source origin.

4.2.2 Historical Dependence and Privacy-Trajectory Attribution

Historical attribution. We examine which historical evidence supports a realized sensitive conclusion. Table 5 shows that Privacy-Trajectory Attribution achieves 0.975 Hit@1, 0.9449 NDCG@3, 0.9060 Support-F1, and 0.803 Exact Support on TOP-Bench. It outperforms Last Turn, Random, and Single-Source attribution on all four metrics.

Table 5: Historical attribution on 1,000 TOP-Bench instances.

Method	Hit@1 ↑	NDCG@3 ↑	Support-F1 ↑	Exact Support ↑
Last Turn	0.000	0.1913	0.3250	0.000
Random	0.562	0.5963	0.5715	0.077
Single-Source	0.954	0.9056	0.8489	0.572
PrivAct (LOO)	**0.975**	**0.9449**	**0.9060**	**0.803**

The Exact Support improvement over Single-Source attribution (0.803 vs. 0.572) shows that isolated relevance is insufficient for historical dependence. On 20 enumerable instances, the full-context LOO estimator also matches exact Shapley at 0.800 Hit@1 while reducing evaluations from 64 to 7 per instance; Appendix B.5 provides the full comparison.

History dependence under StoneSoup. Matched StoneSoup controls further show that disrupting persistent context or the temporal order of evidence reduces attack effectiveness under the same query budget. This result provides attack-side evidence for the context-conditioned historical dependence modeled by Privacy-Trajectory Attribution.

Table 7: Matched protection and runtime comparison on SynthPAI.

Method	Protection Effectiveness (%)					Runtime Efficiency		
	Prec. ↑	F1 ↑	Prevent. ↑	Safe FI ↓	Preserv. ↑	p50 (ms) ↓	p95 (ms) ↓	records/s ↑
No Defense	–	–	0.00	0.00	100.00	–	–	–
DEPN Wu et al. (2023)	26.60	40.65	86.21	14.29	98.79	126.50	342.13	5.95
AirGapAgent Bagdasarian et al. (2024)	24.17	38.93	100.00	17.58	98.61	0.01	0.01	95,706.09
Context Influence Flemings et al. (2025)	30.30	42.11	68.97	14.29	98.92	126.50	342.13	5.95
PrivacyXray He et al. (2025)	50.00	21.62	13.79	2.20	99.83	84.68	298.20	8.40
PrivacyScalpel Frikha et al. (2025)	27.50	40.37	75.86	13.19	98.89	126.50	342.13	5.95
PrivacyChecker Wang et al. (2025)	22.52	35.71	86.21	17.58	98.71	891.47	919.59	1.13
TOP-R Qiao et al. (2025)	19.57	24.00	31.03	15.38	99.19	835.69	865.67	1.30
SumLeaks–ToM Patil et al. (2025)	7.55	9.76	13.79	8.79	99.63	868.37	916.59	1.22
SumLeaks–CoDef Patil et al. (2025)	24.17	38.93	100.00	17.58	98.61	2,341.38	2,702.19	0.43
Stop Tracking Me! Yan et al. (2026)	7.14	4.65	13.79	9.89	91.37	18,070.44	22,100.84	0.06
PATCH Hughes et al. (2026)	100.00	81.63	68.97	0.00	99.48	<0.01	0.02	123,801.03
CDI Wen et al. (2026)	25.25	39.06	86.21	13.19	98.93	891.52	917.46	1.13
OCELOT Xie & Li (2026)	63.04	77.33	100.00	17.58	98.61	589.51	842.56	1.68
PRIVACT (Ours)	**100.00**	**88.46**	**79.31**	**0.00**	**95.15**	**3.84**	**27.13**	**148.85**

Anticipating sensitive disclosure. We finally examine whether historical evidence provides a useful signal before a sensitive claim becomes complete. The future-window prediction experiment shows that the Full Future-Disclosure predictor consistently outperforms Prefix-only prediction across forecasting horizons. At $H = 32$, it improves AUPRC/F1 from 0.60/0.57 to 0.76/0.72 and reduces FPR/ECE from 0.42/0.47 to 0.24/0.27.

The advantage remains consistent as the forecasting horizon increases. Together, these results show that historical dependence serves two complementary roles: it identifies the evidence supporting a sensitive conclusion and provides an early signal before the first-release boundary.

4.2.3 First-Release Protection, Efficiency, and Component Analysis

Cross-model protection. Table 6 shows that PrivAct achieves 79.45–100.00% prevention on open-weight models and 75.00–81.52% on API models. PERMIT FI remains zero, while Task Preservation stays above 99.67%. These measurements indicate protection across the evaluated generators; the 100% Llama-2 prevention estimate is based on just one realized disclosure.

Table 6: Cross-model first-release protection on SynthPAI (%).

Model	Privacy Outcomes			Selectivity / Utility		
	Raw Disc.	Ext. Leak ↓	Prev. ↑	Safe FI ↓	Permit FI ↓	Task Pres. ↑
Baichuan2-7B-Chat	13.67%	14.15%	**85.85%**	3.86%	**0.00%**	**100.00%**
Llama-2-7B-Chat	0.33%	**0.00%**	**100.00%** *	13.04%	**0.00%**	**100.00%**
Mistral-7B-Instruct-v0.2	33.67%	13.86%	**86.14%**	0.50%	**0.00%**	**100.00%**
Qwen2.5-7B-Instruct	24.33%	20.55%	79.45%	0.44%	**0.00%**	**100.00%**
DeepSeek-V4-Flash	27.00%	23.46%	76.54%	**0.00%**	**0.00%**	99.67%
Doubao-Seed-2.0-Mini	24.00%	25.00%	75.00%	**0.00%**	**0.00%**	**100.00%**
GLM-5.2	30.67%	18.48%	**81.52%**	**0.00%**	**0.00%**	**100.00%**

* Based on one realized disclosure.

Matched comparison. Under the same model, samples, and evaluation boundary, PrivAct achieves 100% precision, 88.46% F1, and 0% Safe FI, with a total p50 latency of 3.84 ms (Table 7). The matched comparison also reports 79.31% prevention and 95.15% task preservation, making the tradeoff between selective protection and retained output explicit.

Adaptive StoneSoup protection. Table 8 separates attack success before protection from leakage after protection. Raw ASR measures whether StoneSoup forms a protected disclosure, while External Leakage measures whether it crosses the first-release boundary.

Table 8: Adaptive StoneSoup first-release protection (%).

Model	Raw ASR	Ext. Leak ↓	Prev. ↑	Safe FI ↓	Task Pres. ↑	Warn Before ↑	Lead Time (tokens) ↑
Qwen2.5-7B-Instruct	24.33	20.55	79.45	0.44	100.00	100.00	2
Mistral-7B-Instruct-v0.2	32.10	13.86	86.14	0.50	100.00	100.00	3
Baichuan2-7B-Chat	13.50	14.15	85.85	3.86	100.00	100.00	2
Qwen2.5-7B-Instruct	35.10	20.50	79.50	0.83	100.00	100.00	2
Mistral-7B-Instruct-v0.2	23.70	10.39	**89.61**	0.50	100.00	100.00	3
Baichuan2-7B-Chat	27.50	17.95	82.05	2.18	100.00	100.00	2

Note. The first three rows are evaluated on SynthPAI and the last three on TOP-Bench. External Leakage is conditioned on realized raw disclosures, and Prevention equals $100\% - \texttt{External Leakage}$. Lead Time is the median number of generated tokens between the first valid warning and disclosure completion.

Efficiency and enforcement. Table 9 jointly evaluates end-to-end cost, release behavior, and the necessity of the enforcement chain.

Table 9: Efficiency and enforcement analysis.

(a) End-to-end cost

Model	B	p50 (ms)		Δ Wall	GenTok/s	
		Base	Prot.		Base	Prot.
Llama-3.2-1B	1	658.10	675.01	+2.57%	97.25	94.81
	8	680.56	726.65	+6.77%	752.32	704.61
Gemma-2-2B	1	1445.32	1468.00	+1.57%	44.28	43.60
	8	1477.63	1570.86	+6.31%	346.50	325.94
Qwen2.5-7B	1	1182.29	1197.88	+1.32%	54.13	53.43
	8	1261.05	1310.97	+3.96%	406.01	390.55

(b) Buffer / delay / KV

Model	Buffer		Delay		KV Recomp.
	Mean	Max	Mean	P95	
Baichuan	2.91	31	2.98	8	513
DeepSeek	4.79	36	4.72	12	8
Llama	5.59	39	5.44	16	60
Mistral	2.78	29	3.11	8	1073
Qwen	1.93	32	2.05	4	875

(c) First-release enforcement ablation

Setting	Baichuan2-7B Leak / Prev. / FI	Mistral-7B Leak / Prev. / FI	Qwen2.5-7B Leak / Prev. / FI
Full Chain	14.15 / 85.85 / 3.86	13.86 / 86.14 / 0.50	20.55 / 79.45 / 0.44
− Buffer	100.00 / 0.00 / 0.00	100.00 / 0.00 / 0.00	100.00 / 0.00 / 0.00
− Risk Link	100.00 / 0.00 / 0.00	100.00 / 0.00 / 0.00	100.00 / 0.00 / 0.00
− Local Int.	100.00 / 0.00 / 0.00	100.00 / 0.00 / 0.00	100.00 / 0.00 / 0.00

Each cell reports External Leakage / Prevention / Safe FI (%).

At batch size one, PRIVACT adds only 1.32–2.57% wall-clock latency, with a mean private buffer of 1.93–5.59 tokens. Panel (c) further shows that removing Buffer, Risk Link, or Local Intervention raises External Leakage to 100% and reduces Prevention to 0%. Thus, first-release protection remains lightweight but requires the complete enforcement chain.

5 DISCUSSION AND CONCLUSION

Scope and limitations. The experiments use public benchmark instances and a selected set of generators; their prevention rates should not be read as guarantees for unseen deployments. Historical leave-one-out attribution measures counterfactual model support conditional on a fixed target claim and observed history, not a causal effect or an online explanation available before the claim is formed. The future-window predictor is the distinct online warning mechanism. Prevention also depends on the deployment policy, claim recognition, and the ability to keep candidate output private until a decision is made. Longer persistent-memory, retrieval, and agentic histories merit separate validation, as do stronger adaptive elicitation strategies.

Deployment and responsible use. An operator can enforce a conservative release boundary by buffering text until policy checks complete and by holding content when the risk threshold cannot be met. A safe rewrite can preserve useful output, but its quality and authorization must be checked under the application's own policy. Our experiments use the described privacy benchmarks; claims about real user privacy or vendor remediation require separate evidence. These constraints make the release boundary, policy ownership, and treatment of retained conversation histories important deployment decisions.

Conclusion. PRIVACT treats sensitive disclosure in multi-turn LLM interactions as a claim that may form from distributed historical evidence and then cross an external release boundary. It separates source-explicit disclosures from Novel Sensitive Claims, reconstructs historical support through a Privacy Trajectory, and combines online warning with selective buffered intervention. On the evaluated benchmarks, attribution reaches 0.803 Exact Support on TOP-Bench; the matched protection comparison yields 100% detection precision and 79.31% prevention. These findings motivate privacy checks that consider both how evidence accumulates and when model output becomes observable.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, literature retrieval and discovery, and code development and debugging. In particular, these tools were used to improve grammar, clarity, and presentation, to help identify potentially relevant prior work, and to assist with software implementation. All retrieved references were manually verified against original publisher pages, proceedings records, or official repository pages before being included in the paper. All AI-assisted code was manually reviewed and tested by the authors. The authors independently reviewed all AI-assisted edits and verified the technical content, experimental results, mathematical statements, and citations. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies privacy risks and defenses for multi-turn LLM interactions and therefore has potential dual-use implications. The StoneSoup attack is introduced to evaluate whether an unauthorized requester can exploit interaction history to induce protected disclosures. Our experiments use public synthetic or benchmark data and do not involve human subjects or real private user records. We report the attack to support systematic evaluation and mitigation rather than deployment against real users. PRIVACT is intended as a research framework for measuring and controlling privacy risk before external release. The reported results should not be interpreted as guarantees for real-world deployment. Practical use requires application-specific privacy policies, authorization rules, secure handling of conversation history, and appropriate controls on retained candidate content. Any deployment involving real user histories should additionally satisfy applicable consent, privacy, security, and data-governance requirements.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce the reported experiments across the main paper and appendix. Section 3 defines the claim-level privacy semantics, Privacy-Trajectory Attribution, future-window risk measurement, and pre-release intervention mechanism. Section 4 reports the primary evaluation protocol and results. Appendix A specifies the benchmarks, model settings, NSC evaluation procedure, Adaptive StoneSoup protocol, baseline adaptation, metric definitions, runtime boundaries, and reproducibility settings. Appendix B provides the full matched comparisons, detailed runtime results, additional StoneSoup results, intervention behavior, and the comparison with exact Shapley attribution. All tunable thresholds are selected before test evaluation and are held fixed during testing. Exact model identifiers and evaluation settings are reported where applicable.

REFERENCES

Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. AirGapAgent: Protecting privacy-conscious conversational agents. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3868–3882. ACM, 2024. doi: 10.1145/3658644.3690350. URL https://doi.org/10.1145/3658644.3690350.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Junjie Chu, Zeyang Sha, Michael Backes, and Yang Zhang. Reconstruct your previous conversations! comprehensively investigating privacy leakage risks in conversations with GPT models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6584–6600, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.377. URL https://aclanthology.org/2024.emnlp-main.377/.

Dave Citron. Reference past chats for more tailored help with Gemini Advanced. Google Blog, February 2025. URL https://blog.google/feed/gemini-referencing-past-chats/. Published February 13, 2025; accessed September 25, 2026.

Saswat Das and Ferdinando Fioretto. NeuroFilter: Privacy guardrails for conversational LLM agents, 2026. URL https://arxiv.org/abs/2601.14660.

Saswat Das, Jameson Sandler, and Ferdinando Fioretto. Disclosure audits for LLM agents, 2025. URL https://arxiv.org/abs/2506.10171. Earlier versions circulated under the title "Beyond Jailbreaking: Auditing Contextual Privacy in LLM Agents".

DeepSeek. DeepSeek-V4-Flash update. DeepSeek API Documentation, July 2026. URL https://api-docs.deepseek.com/updates/. Official V4-Flash API release dated July 31, 2026; accessed September 25, 2026.

Charles Duhigg. How companies learn your secrets. *The New York Times Magazine*, February 2012. URL https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html. Published February 16, 2012; accessed September 4, 2026.

Faouzi El Yagoubi, Godwin Badu-Marfo, and Ranwa Al Mallah. AgentLeak: A benchmark for internal-channel privacy leakage in multi-agent LLM systems. *IEEE Access*, 14:94960–94978, 2026. doi: 10.1109/ACCESS.2026.3704541. URL https://ieeexplore.ieee.org/document/11569042.

James Flemings, Bo Jiang, Wanrong Zhang, Zafar Takhirov, and Murali Annavaram. Estimating privacy leakage of augmented contextual knowledge in language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25092–25108, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1220. URL https://aclanthology.org/2025.acl-long.1220/.

Ahmed Frikha, Muhammad Reza Ar Razi, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. PrivacyScalpel: Enhancing LLM privacy via interpretable feature intervention with sparse autoencoders. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 226–238, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.blackboxnlp-1.13. URL https://aclanthology.org/2025.blackboxnlp-1.13/.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Leonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

GLM-5 Team et al. GLM-5: From vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Chi Han, Xin Liu, Haodong Wang, Shiyang Li, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Qingyu Yin, Liang Qiu, Changlong Yu, Yifan Gao, Zheng Li, Bing Yin, Jingbo Shang, and Heng Ji. Can language models follow multiple turns of entangled instructions? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 25445–25460, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1387. URL https://aclanthology.org/2025.findings-emnlp.1387/.

Jinwen He, Yiyang Lu, Zijin Lin, Kai Chen, and Yue Zhao. PrivacyXray: Detecting privacy breaches in LLMs through semantic consistency and probability certainty. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 8195–8214, Seattle, WA, 2025. USENIX Association. URL https://www.usenix.org/conference/usenixsecurity25/presentation/he-jinwen.

Anthony Hughes, Vasisht Duddu, N. Asokan, Nikolaos Aletras, and Ning Ma. PATCH: Mitigating PII leakage in language models with privacy-aware targeted circuit PatcHing. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5139–5153, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-eacl.271. URL https://aclanthology.org/2026.findings-eacl.271/.

Information Commissioner's Office. What is special category data? Information Commissioner's Office, April 2024. URL https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-is-special-category-data/. Guidance updated April 9, 2024; accessed September 25, 2026.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 346–363. IEEE, 2023. doi: 10.1109/SP46215.2023.10179300.

Meta AI. Llama 3.2: Model cards and prompt formats. Meta AI, 2024. URL https://dev.meta.ai/llama/docs/model-cards-and-prompt-formats/llama3_2. Official documentation for Llama 3.2 lightweight 1B/3B models; accessed September 25, 2026.

OpenAI. Memory and new controls for ChatGPT. OpenAI, February 2024. URL https://openai.com/index/memory-and-new-controls-for-chatgpt/. Published February 13, 2024; updated June 3, 2025; accessed September 25, 2026.

Vaidehi Patil, Elias Stengel-Eskin, and Mohit Bansal. The sum leaks more than its parts: Compositional privacy risks and mitigations in multi-agent collaboration, 2025. URL https://arxiv.org/abs/2509.14284.

Yuxuan Qiao, Dongqin Liu, Hongchang Yang, Wei Zhou, and Songlin Hu. Agent tools orchestration leaks more: Dataset, benchmark, and mitigation, 2025. URL https://arxiv.org/abs/2512.16310.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9028b8a3ca98f58e373f0c1497a17448-Abstract-Conference.html.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Language models are advanced anonymizers. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f478b2e8ad9ff0756bf5b79fb31c330f-Abstract-Conference.html.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Volcano Engine. Doubao Seed 2.0 Mini. Volcano Engine Ark Documentation, 2026. URL https://docs.volcengine.com/docs/ark/model-list?lang=zh. Model ID: doubao-seed-2-0-mini-260428; accessed September 25, 2026.

Shouju Wang, Fenglin Yu, Xirui Liu, Xiaoting Qin, Jue Zhang, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. Privacy in action: Towards realistic privacy mitigation and evaluation for LLM-powered agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17055–17074, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.925. URL https://aclanthology.org/2025.findings-emnlp.925/.

Yule Wen, Yanzhe Zhang, Jianxun Lian, Xiaoyuan Yi, Xing Xie, and Diyi Yang. Contextualized privacy defense for LLM agents, 2026. URL https://arxiv.org/abs/2603.02983.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. DEPN: Detecting and editing privacy neurons in pretrained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2875–2886, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.174. URL https://aclanthology.org/2023.emnlp-main.174/.

Xinwei Wu, Weilong Dong, Shaoyang Xu, and Deyi Xiong. Mitigating privacy seesaw in large language models: Augmented privacy neuron editing via activation patching. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5319–5332, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.315. URL https://aclanthology.org/2024.findings-acl.315/.

Jin Xie and Songze Li. OCELOT: Inference-leakage budgets for privacy-preserving LLM agents, 2026. URL https://arxiv.org/abs/2606.12341.

Dong Yan, Jian Liang, Ran He, and Tieniu Tan. Stop tracking me! proactive defense against attribute inference attack in LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f1c131d41c3d5119a8de329140b3384e-Abstract-Conference.html.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models, 2023. URL https://arxiv.org/abs/2309.10305.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. Robust utility-preserving text anonymization based on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28922–28941, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1404. URL https://aclanthology.org/2025.acl-long.1404/.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. A synthetic dataset for personal attribute inference. In *Advances in Neural Information Processing Systems*, volume 37, pp. 120735–120779. Neural Information Processing Systems Foundation, Inc., 2024. doi: 10.52202/079017-3837. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/daa1816b84ca2d5051c87fb4d37dd540-Abstract-Datasets_and_Benchmarks_Track.html.

A Experimental Details

This section provides the experimental details omitted from the main paper. We describe the evaluation instances, model settings, NSC evaluation, StoneSoup protocol, matched baseline adaptation, and runtime measurement. All tunable thresholds are selected before test evaluation and remain fixed throughout testing.

A.1 Benchmarks and Evaluation Instances

We primarily use SynthPAI Yukhymenko et al. (2024) and TOP-Bench Qiao et al. (2025). SynthPAI is a public synthetic benchmark for personal attribute inference. Its public release contains synthetic user profiles and Reddit-style comments generated by profile-conditioned LLM agents. We use the 300 public profiles and their associated contextual records to construct the interaction histories used in our privacy-formation, StoneSoup, cross-model protection, and matched comparison experiments. TOP-Bench contains 1,000 evaluation cases for privacy risks that emerge from combining information across multiple tool results. Each case provides a target sensitive conclusion and a `minimal_supporting_set` that identifies the evidence jointly needed to support that conclusion. We use these supporting sets as reference evidence for historical attribution. For

Table 10: Models and their roles in the evaluation.

Model	Access	Evaluation Role
Qwen2.5-7B-Instruct	Local	Primary, matched comparison, StoneSoup
Baichuan2-7B-Chat	Local	Formation, protection, StoneSoup
Llama-2-7B-Chat	Local	Formation, cross-model protection
Mistral-7B-Instruct-v0.2	Local	Formation, protection, StoneSoup
DeepSeek-V4-Flash	API	Cross-model protection
Doubao-Seed-2.0-Mini	API	Cross-model protection
GLM-5.2	API	Cross-model protection
Llama-3.2-1B	Local	End-to-end efficiency
Gemma-2-2B	Local	End-to-end efficiency

Table 11: Representative outcomes under the NSC formulation.

Representative Claim Condition	Decision	Reason
A protected sensitive fact is already explicit in a historical source	Source-explicit	The information already appears in the available history.
Multiple historical units jointly support a new protected sensitive conclusion	NSC	The full NSC criterion is satisfied, including source novelty.
A source-novel sensitive claim fails the remaining NSC conditions	Non-NSC	Source novelty alone is insufficient.
The generated claim does not satisfy the protected privacy semantics	Non-sensitive	The claim falls outside the protected claim space.

the history-dependence experiments, we treat the corresponding tool-result units as temporally ordered historical evidence. This use preserves the benchmark's compositional privacy structure while adapting it to our multi-turn attribution setting.

A.2 Models and Generation Settings

We use Qwen2.5-7B-Instruct Yang et al. (2024) as the primary generator. The open-weight cross-model evaluation further includes Baichuan2-7B-Chat Yang et al. (2023), Llama-2-7B-Chat Touvron et al. (2023), and Mistral-7B-Instruct-v0.2 Jiang et al. (2023). Efficiency experiments additionally use Llama-3.2-1B Meta AI (2024) and Gemma-2-2B Gemma Team et al. (2024). We also evaluate the API models DeepSeek-V4-Flash DeepSeek (2026), Doubao-Seed-2.0-Mini Volcano Engine (2026), and GLM-5.2 GLM-5 Team et al. (2026).

Within each matched comparison, we keep the base generator, evaluation samples, and generation boundary fixed. Different methods only change the privacy-checking or intervention path. Within each cross-model experiment, we use the same task prompts and keep the model-specific generation configuration fixed across compared conditions.

A.3 NSC Recognition and Evaluation

PrivAct applies the same NSC criterion defined in the main method at runtime. For each candidate claim c, the decision depends on the claim, the available historical sources D, the current request q, and the privacy policy Π. In particular, source novelty is necessary but not sufficient for an NSC. A source-novel claim that does not satisfy the remaining claim-level privacy conditions is not classified as an NSC. Likewise, a protected sensitive claim that is already explicit in a historical source is treated as source-explicit rather than source-novel.

A.4 StoneSoup Attack Protocol

A StoneSoup session fixes a target subject u, a protected attribute a, an initial history D_0, and a maximum query budget T. At turn t, Adaptive StoneSoup selects the next query q_t from the previously observed interaction history H_{t-1}. The target model receives the current context $D_t = \mathrm{Concat}(D_0, H_{t-1})$. The attacker does not need to know or explicitly state the target sensitive value. Instead, successive queries progressively activate and connect historical evidence related to the protected attribute.

Algorithm 1 Adaptive StoneSoup Attack

Require: subject u, protected attribute a, initial history D_0, model M, policy Π, budget T

$H_0 \leftarrow \emptyset$
for $t = 1, \ldots, T$ **do**
  $q_t \leftarrow \mathcal{A}(u, a, D_0, H_{t-1})$
  $D_t \leftarrow \mathrm{Concat}(D_0, H_{t-1})$
  $y_t \leftarrow M(D_t, q_t)$
  $\mathcal{C}_t \leftarrow \mathrm{ExtractClaims}(y_t)$
  for all $c \in \mathcal{C}_t$ **do**
    if PROTECTEDDISCLOSURE(c, D_t, q_t, Π) **then**
      return $(\mathsf{Success}, t, c)$
    end if
  end for
  $H_t \leftarrow H_{t-1} \parallel (q_t, y_t)$
end for
return $(\mathsf{Failure}, T, \emptyset)$

Table 12: Matched interfaces used for baseline comparison.

Methods	Interface	Matched Evaluation Form
No Defense	ER	External release
DEPN, Context Influence, PrivacyXray, PrivacyScalpel	D+LG	Detection + local gate
AirGapAgent, PrivacyChecker, TOP-R, SumLeaks, CDI, OCELOT	OG	Output gate
Stop Tracking Me!	P+R	Preprocessing + regeneration
PATCH	EL+LG	Exact localization + local gate
PRIVACT	BFR	Before-first-release protection

ER = external release; D+LG = detection plus local gate; OG = output gate; P+R = preprocessing plus regeneration; EL+LG = exact localization plus local gate; BFR = before first release.

For source-novel evaluation, an attack succeeds when a generated claim first satisfies the NSC criterion $N(c; D_t, q_t, \Pi) = 1$ defined in the main method. For source-explicit evaluation, we use the same interaction protocol but apply the corresponding protected source-explicit disclosure criterion.

We compare Adaptive StoneSoup with four controls. Direct Query uses a single request. Fixed Multi-turn follows a predetermined query sequence. Shuffled Multi-turn uses the same query set in a different order. In Reset-History StoneSoup, the attacker retains previously observed responses, but the target model receives no preceding conversational turns at the next query. All multi-turn settings use the same maximum query budget.

For source-novel sessions, PROTECTEDDISCLOSURE uses $N(c; D_t, q_t, \Pi) = 1$. For source-explicit sessions, it uses the corresponding source-explicit privacy criterion. The same adaptive interaction process is therefore used for both source-origin categories.

A.5 BASELINE ADAPTATION AND METRIC DEFINITIONS

The compared methods were originally developed for different privacy objectives and deployment interfaces. We therefore use a matched protocol that fixes the evaluation samples, base generator, and release boundary. Where the original interface permits, we preserve each method's primary decision signal and adapt only the surrounding deployment interface required for matched evaluation. The compact effectiveness–efficiency comparison is reported in the main paper. Appendix B provides the complete detection, intervention, and runtime breakdowns.

Raw Disclosure and Raw ASR are measured before protection. External Leakage is measured only after the protection mechanism is applied. This separation distinguishes privacy formation or attack success from defense failure.

A.6 RUNTIME AND REPRODUCIBILITY SETTINGS

We use two timing boundaries. Matched runtime excludes base-model generation and measures only the method-specific privacy-checking and intervention path. All methods are evaluated on the same

Table 13: Definitions and evaluation populations of the main metrics.

Metric	Definition	Population
Raw Disclosure	Protected disclosure produced before intervention.	Evaluated generations
Raw ASR	Attack session that produces the target protected disclosure before intervention.	Attack sessions
NSC ASR	Attack session that produces a successful NSC.	Attack sessions
External Leakage	Realized protected disclosure that crosses the external release boundary.	Realized disclosures
Prevention	Realized protected disclosure blocked before external release.	Realized disclosures
Safe FI	Intervention applied to an output that does not require protection.	Safe outputs
PERMIT FI	Intervention applied when the policy explicitly permits the disclosure.	PERMIT outputs
Task Preservation	Protected output that preserves the requested task objective.	Evaluated task outputs
Hit@1	Whether the highest-ranked historical unit belongs to the reference support set.	Attribution cases
NDCG@3	Ranking quality of the top attributed historical units.	Attribution cases
Support-F1	Set overlap between predicted and reference historical support.	Attribution cases
Exact Support	Exact match between predicted and reference support sets.	Attribution cases
AUPRC / F1	Discrimination quality of Future-Disclosure prediction.	Future-window cases
FPR / ECE	False-positive behavior and calibration of Future-Disclosure prediction.	Future-window cases

Table 14: Runtime and reproducibility settings used in the reported experiments.

Setting	Configuration
Matched benchmark	SynthPAI
Matched samples	1200 outputs
Matched base model	Qwen2.5-7B-Instruct
Matched timing boundary	Privacy checking + intervention only
Runtime statistics	p50 / p95 / records/s
End-to-end generation	64 tokens
Batch sizes	$B = 1, 8$
Efficiency models	Llama-3.2-1B, Gemma-2-2B, Qwen2.5-7B
Buffer statistics	Mean / Max
Release-delay statistics	Mean / P95

1200 SynthPAI outputs with the same base model, batch setting, hardware, and timing boundary. The main paper reports total p50, total p95, and throughput. Appendix B provides the complete checking and intervention breakdown.

End-to-end runtime measures the complete protected generation process. We use 64-token generation with batch sizes $B = 1$ and $B = 8$. We additionally record private-buffer size and release delay to characterize the cost of enforcing the first-release boundary.

Matched runtime and end-to-end runtime therefore measure different costs. We report them separately and do not interpret method-level checking latency as complete generation latency.

B Additional Results and Analysis

This section provides additional results that complement the compact comparisons in the main paper. We report the full matched protection results, the detailed runtime breakdown, additional StoneSoup robustness experiments, intervention behavior, and a direct comparison between Full-Context LOO and exact Shapley attribution. Results already reported in full in the main paper are not repeated here.

B.1 Full Matched Baseline Results

The compact comparison in the main paper reports only total p50, total p95, and throughput. Here, we further decompose the runtime into privacy checking and intervention. All methods process the same SynthPAI outputs under the same base model, batch size, hardware, and timing boundary. This matched measurement isolates the privacy-protection path and excludes base-model generation.

For PrivAct, privacy checking requires 3.52 ms at p50 and 15.76 ms at p95, while intervention requires 0.28 ms at p50 and 11.45 ms at p95. Direct measurement of the complete protection path

Table 15: Full matched comparison of sensitive-leakage detection and first-release intervention on SynthPAI using Qwen2.5–7B–Instruct.

Method	Sensitive-Leakage Detection (%)				First-Release Intervention				
	Acc. ↑	Prec. ↑	Recall ↑	F1 ↑	Interface	Residual ↓	Prevent. ↑	Safe FI ↓	Preserv. ↑
No Defense	–	–	–	–	ER	100.00%	0.00%	0.00%	100.00%
DEPN Wu et al. (2023)	39.17	26.60	86.21	40.65	D+LG	13.79%	86.21%	14.29%	98.79%
AirGapAgent Bagdasarian et al. (2024)	24.17	24.17	100.00	38.93	OG	0.00%	100.00%	17.58%	98.61%
Context Influence Flemings et al. (2025)	54.17	30.30	68.97	42.11	D+LG	31.03%	68.97%	14.29%	98.92%
PrivacyXray He et al. (2025)	75.83	50.00	13.79	21.62	D+LG	86.21%	13.79%	2.20%	99.83%
PrivacyScalpel Frikha et al. (2025)	45.83	27.50	75.86	40.37	D+LG	24.14%	75.86%	13.19%	98.89%
PrivacyChecker Wang et al. (2025)	25.00	22.52	86.21	35.71	OG	13.79%	86.21%	17.58%	98.71%
TOP-R Qiao et al. (2025)	52.50	19.57	31.03	24.00	OG	68.97%	31.03%	15.38%	99.19%
SumLeaks–ToM Patil et al. (2025)	38.33	7.55	13.79	9.76	OG	86.21%	13.79%	8.79%	99.63%
SumLeaks–CoDef Patil et al. (2025)	24.17	24.17	100.00	38.93	OG	0.00%	100.00%	17.58%	98.61%
Stop Tracking Me! Yan et al. (2026)	65.83	7.14	3.45	4.65	P+R	86.21%	13.79%	9.89%	91.37%
PATCH Hughes et al. (2026)	92.50	100.00	68.97	81.63	EL+LG	31.03%	68.97%	0.00%	99.48%
CDI Wen et al. (2026)	35.00	25.25	86.21	39.06	OG	13.79%	86.21%	13.19%	98.93%
OCELOT Xie & Li (2026)	85.83	63.04	100.00	77.33	OG	0.00%	100.00%	17.58%	98.61%
PRIVACT (Ours)	**95.00**	**100.00**	**79.31**	**88.46**	**BFR**	**20.69%**	**79.31%**	**0.00%**	**95.15%**

Interface codes: ER = external release; D+LG = detection plus local gate; OG = output gate; P+R = preprocessing plus regeneration; EL+LG = exact localization plus local gate; BFR = before first release.

Table 16: Detailed mechanism-level runtime of all methods on the same SynthPAI outputs under matched base model, batch size, hardware, and timing boundaries.

Method	Check p50 (ms) ↓	Check p95 (ms) ↓	Interv. p50 (ms) ↓	Interv. p95 (ms) ↓	Total p50 (ms) ↓	Total p95 (ms) ↓	records/s ↑
DEPN Wu et al. (2023)	126.50	342.12	0.01	0.01	126.50	342.13	5.95
AirGapAgent Bagdasarian et al. (2024)	<0.01	<0.01	0.01	0.01	0.01	0.01	95,706.09
Context Influence Flemings et al. (2025)	126.50	342.12	0.00	0.01	126.50	342.13	5.95
PrivacyXray He et al. (2025)	84.68	298.20	0.00	0.01	84.68	298.20	8.40
PrivacyScalpel Frikha et al. (2025)	126.50	342.12	0.01	0.01	126.50	342.13	5.95
PrivacyChecker Wang et al. (2025)	891.46	919.58	0.01	0.01	891.47	919.59	1.13
TOP-R Qiao et al. (2025)	835.68	865.66	0.01	0.01	835.69	865.67	1.30
SumLeaks–ToM Patil et al. (2025)	868.36	916.58	0.01	0.01	868.37	916.59	1.22
SumLeaks–CoDef Patil et al. (2025)	2,341.38	2,702.18	0.01	0.01	2,341.38	2,702.19	0.43
Stop Tracking Me! Yan et al. (2026)	17,329.11	21,258.97	755.04	1,012.07	18,070.44	22,100.84	0.06
PATCH Hughes et al. (2026)	<0.01	<0.01	0.00	0.01	<0.01	0.02	123,801.03
CDI Wen et al. (2026)	891.51	917.45	0.01	0.01	891.52	917.46	1.13
OCELOT Xie & Li (2026)	589.50	842.55	0.01	0.01	589.51	842.56	1.68
PRIVACT (Ours)	**3.52**	**15.76**	**0.28**	**11.45**	**3.84**	**27.13**	**148.85**

Note. Matched runtime excludes base-model generation and measures only the method-specific privacy-checking and intervention path.

yields a total latency of 3.84 ms at p50 and 27.13 ms at p95, with a throughput of 148.85 records/s. This decomposition shows that the median runtime is dominated by privacy checking, while local intervention adds only limited overhead in the typical case. Because matched runtime measures only the protection path, we separately report end-to-end generation overhead in the main paper. At batch size one, PRIVACT increases wall-clock latency by only 1.32–2.57%.

B.2 Detailed Matched Runtime Breakdown

The compact comparison in the main paper reports only total latency and throughput. Table 16 further separates privacy checking from intervention. All methods process the same SynthPAI outputs under the same base model, batch size, hardware, and timing boundary. Base-model generation is excluded from this matched measurement.

For PRIVACT, privacy checking requires 3.52 ms at p50 and 15.76 ms at p95. Local intervention adds 0.28 ms at p50 and 11.45 ms at p95. The resulting total latency is 3.84 ms at p50 and 27.13 ms at p95. Some lightweight output gates are faster under this matched interface. However, these measurements represent only the adapted protection path and should not be interpreted as the complete cost of each original system. The end-to-end measurements in the main paper separately capture the full cost of protected generation.

B.3 Additional StoneSoup Robustness

The main paper shows that Adaptive StoneSoup can actively exploit multi-turn privacy formation. We further examine whether PrivAct can still prevent a sensitive conclusion from crossing the first-release boundary after the attacker successfully exploits the interaction history.

Table 17 shows that Raw ASR varies substantially across models and benchmarks. Nevertheless, PrivAct prevents 79.45–89.61% of realized disclosures, while Safe FI remains at or below 3.86% and Task Preservation remains 100% in all evaluated settings. All evaluated disclosures receive a valid warning before completion, with a median lead time of 2–3 generated tokens. These results show that first-release protection does not depend on a fixed StoneSoup query template. Instead, PrivAct evaluates the current claim against its historical evidence and intervenes before the protected disclosure becomes externally visible.

Table 17: Adaptive StoneSoup first-release protection (%).

Model	Raw ASR	Ext. Leak ↓	Prev. ↑	Safe FI ↓	Task Pres. ↑	Warn Before ↑	Lead Time (tokens) ↑
Qwen2.5-7B-Instruct	24.33	20.55	79.45	0.44	100.00	100.00	2
Mistral-7B-Instruct-v0.2	32.10	13.86	86.14	0.50	100.00	100.00	3
Baichuan2-7B-Chat	13.50	14.15	85.85	3.86	100.00	100.00	2
Qwen2.5-7B-Instruct	35.10	20.50	79.50	0.83	100.00	100.00	2
Mistral-7B-Instruct-v0.2	23.70	10.39	**89.61**	0.50	100.00	100.00	3
Baichuan2-7B-Chat	27.50	17.95	82.05	2.18	100.00	100.00	2

Note. The first three rows report SynthPAI results and the last three report TOP-Bench results. External Leakage is conditioned on realized raw disclosures, and Prevention equals $100\% - \texttt{External Leakage}$. Warn Before reports the fraction of realized disclosures with a valid warning before completion. Lead Time is the median number of generated tokens between the first valid warning and disclosure completion.

B.4 Output Preservation and Intervention Behavior

Table 18 examines whether first-release intervention preserves task-relevant content across models. PrivAct maintains high semantic similarity on all four open-weight models, with BERTScore F1 ranging from 91.04% to 95.30%. The Win Rate identification intervals also remain high across models. Rewrite Safe reaches 100% on Llama, Mistral, and Qwen and remains between 98.02% and 100.00% on Baichuan.

These results support the design of Minimal-Risk Pre-release Intervention. PrivAct does not map every detected risk to a global refusal. Instead, it selects among Continue, Rewrite, and Hold according to the privacy state of the current generation. The high semantic similarity shows that protected outputs preserve most of their task-relevant content, while the high Rewrite Safe rate shows that local rewriting can often remove the protected disclosure without discarding the entire response.

Table 18: Output preservation and intervention behavior across open-weight models.

Model	BERTScore F1 (%) ↑	Win Rate (%) ↑	Continue (%)	Rewrite (%)	Hold (%)	Rewrite Safe (%) ↑
Baichuan2-7B	91.304	[91.71, 92.04]	17.83	38.42	43.75	[98.02, 100.00]
Llama-2-7B	**95.304**	[90.29, 94.54]	28.25	30.50	41.25	100.00
Mistral-7B	92.265	[90.29, 92.54]	13.15	42.50	44.35	100.00
Qwen2.5-7B	91.043	[90.42, 95.00]	13.50	43.25	53.25	100.00

Note. BERTScore F1 uses the frozen semantic-similarity backbone. Win Rate intervals preserve unresolved pairwise judgments and are identification intervals, not confidence intervals. Rewrite Safe is computed only over outputs for which PrivAct selects Rewrite.

B.5 Attribution Fidelity and Computational Cost

Privacy-Trajectory Attribution uses a full-context leave-one-out estimator to measure the contribution of each historical unit to a fixed sensitive claim. Exact Shapley provides a more exhaustive coalition-based attribution, but its evaluation cost grows rapidly with the number of historical units. We therefore compare the two approaches on 20 instances for which all relevant coalitions can be enumerated exactly.

As shown in Table 19, both methods achieve 0.800 Hit@1 on this subset. Exact Shapley requires 64 model evaluations per instance, whereas Privacy-Trajectory Attribution requires only seven. Thus,

the proposed estimator achieves the same aggregate Hit@1 at substantially lower evaluation cost. This result supports its use for runtime historical attribution.

Table 19: Attribution fidelity and evaluation cost on 20 enumerable instances.

Method	**Instances**	**Hit@1** ↑	**Evaluations / Instance** ↓
Exact Shapley	20	0.800	64
Privacy-Trajectory Attribution (Ours)	20	**0.800**	**7**