

GuardianAgent: Policy-Conditioned Risk-Adaptive Anonymization with Verified Adversarial Escalation

Ruiyi Yang¹ Gayathri Lihinikaduarachchi¹ Rahat Masood¹
 Flora D. Salim¹ Salil S. Kanhere¹

¹School of Computer Science and Engineering, UNSW Sydney, Australia
 {ruiyi.yang, g.lihinikadu_arachchillage}@student.unsw.edu.au
 {rahat.masood, flora.salim, salil.kanhere}@unsw.edu.au

Abstract

Privacy protection for live web traffic requires more than detecting private spans. Agent-based privacy protection systems must determine whether an outgoing action complies with the destination site’s privacy policy, then apply only the level of rewriting or sanitisation justified by the residual disclosure risk. We present **GuardianAgent**, a policy-conditioned anonymization framework that couples structured risk assessment with verified adaptive rewriting. GuardianAgent computes risk through **AMRSF** (Adaptive Multi-factor Risk Scoring Formula), an explicit controller that combines policy-violation likelihood with data sensitivity, recipient transmission, purpose legitimacy, contextual basis, and policy transparency, rather than relying on an LLM to assign risk directly. This risk score determines both the allow/transform/deny decision and the initial anonymization level. For efficiency, GuardianAgent uses an evidential fast path for low-uncertainty policy matches and invokes an LLM slow path only for uncertain cases. For rewriting, it applies a **five-level hierarchy** driven by a **verified adversarial guesser**: guesses trigger escalation only when supported by the original text, preventing hallucinated attacker confidence from causing unnecessary over-anonymization. Experiments across three benchmarks spanning legal text (TAB), Reddit posts (SynthPAI), and multi-format synthetic PII records (PII-Masking-300k) show that GuardianAgent achieves the strongest privacy–utility trade-off among published baselines and is the only method to reach ≥ 0.90 privacy in all three domains, remaining robust under a backbone switch. Action-context stress tests further show that the same outgoing text receives different decisions and anonymization strengths under different recipients, purposes, action bases, and policy-transparency conditions.¹

¹Code and evaluation harness at <https://anonymous.4open.science/r/guardianAgent/>.

1 Introduction

Modern web applications routinely expose fine-grained personal signals through form fields, paste events, outbound requests, clipboard content, and chatbot interactions. A browser-based *privacy agent* should therefore intervene before an action leaves the browser: it must determine whether the action is consistent with the destination site’s privacy policy, and when needed, rewrite the outgoing content to reduce privacy leakage (Zhou et al., 2025; Staab et al., 2025; Chen et al., 2023b; Mireshghallah et al., 2023). This setting brings together two NLP problems that are usually studied in isolation: **policy matching**, which grounds an observed action in a heterogeneous privacy policy, and **text anonymization**, which rewrites content so that sensitive attributes cannot be inferred.

Existing work leaves three gaps. First, policy-understanding systems classify, retrieve, or summarize privacy-policy statements (Harkous et al., 2018; Ahmad et al., 2021; Srinath et al., 2021a; Tang et al., 2023; Sun et al., 2025), and recent LLM-based interfaces make policies easier to query and assess (Chen et al., 2025; Freiburger et al., 2026). However, these systems do not provide an operational risk signal that links live action context to the required strength of anonymization. The same text may be acceptable with light rewriting when sent to a first-party service for core functionality, but require stronger rewriting or blocking when sent to an advertising network for profiling. Second, recent LLM-based anonymizers often invoke a large model for every browser event (Staab et al., 2025; Shao et al., 2025), making them poorly suited for low-latency browser-resident mediation. Third, adversarial anonymizers commonly escalate rewriting whenever an auxiliary attacker reports high confidence. Such confidence, however, may reflect hallucinated priors rather than genuine residual leakage. For example, given [PERSON] going

to [LOCATION] this afternoon, an attacker may confidently guess “coffee shop” although the original was “school with Bob”, causing unnecessary over-anonymization.

This paper presents **GuardianAgent**, a policy-conditioned privacy agent that couples structured risk assessment with verified adaptive anonymization. GuardianAgent computes risk through AMRSF, an explicit controller that combines policy-violation likelihood with data sensitivity, recipient transmission, purpose legitimacy, contextual basis, and policy transparency. The resulting score jointly determines the allow/transform/deny decision and the initial anonymization level, ensuring that rewriting strength is controlled by action context rather than by an LLM-assigned risk judgment alone. To reduce latency, GuardianAgent uses an evidential fast path for low-uncertainty policy matches and routes only uncertain cases to an LLM slow path. The anonymizer then applies a five-level rewrite hierarchy and escalates only when a verified adversarial guesser identifies residual leakage that is actually supported by the original text. Our main contributions are:

- **Policy-conditioned risk-adaptive anonymization:** we introduce a structured risk score that jointly governs action-level decisions and anonymization strength. Action-context stress tests show that the same outgoing text receives different decisions and initial anonymization levels depending on recipient, purpose, action-basis, and policy transparency.
- **Verified adversarial anonymization:** we propose a hierarchical anonymizer that escalates rewriting only when adversarial guesses are supported by the original text, preventing hallucinated confidence from triggering unnecessary semantic distortion.
- **Cross-domain empirical validation:** experiments on legal text (TAB), Reddit posts (SynthPAI), and multi-format PII records (PII-Masking-300k) show that GuardianAgent consistently improves the privacy–utility trade-off over published baselines, is the only method to reach ≥ 0.90 privacy across all three domains, and maintains its gains under a back-bone switch.

2 Related Work

Privacy-policy grounding. Prior work has developed methods for extracting structured statements from privacy policies to enable downstream systems to reason about data practices. Polisris (Harkous et al., 2018) trained classifiers on the OPP-115 schema (Wilson et al., 2016); PrivaSeer (Srinath et al., 2021a) built a corpus-scale search infrastructure for privacy policies; PrivBERT (Srinath et al., 2021b) and PolicyGPT (Tang et al., 2023) studied domain-adaptive and LLM-based policy understanding; and Ahmad et al. (2021) framed policy analysis as joint intent classification and slot filling. These systems make privacy policies more machine-interpretable, but they primarily support descriptive analysis: they do not directly ground a live browser action in the relevant policy evidence, nor do they determine whether the outgoing content should be allowed, rewritten, or blocked.

Privacy-risk frameworks. Quantitative privacy-risk modelling draws on regulatory tiering and contextual theories, including NIST SP 800-122 (McCallister, 2010) and GDPR special categories (Regulation, 2016), FAIR-style likelihood–impact products (Cronk and Shapiro, 2021), and contextual integrity (Nissenbaum, 2009; Miresghallah et al., 2023). Recent LLM-based systems further support privacy-policy assessment and risk comprehension (Sun et al., 2025; Xie et al., 2025), while agent-oriented work studies privacy risks in interactive LLM systems (Ngong et al., 2025). These works are mainly used for explanation, auditing, or evaluation; GuardianAgent instead uses risk factors as a runtime controller linking policy evidence and action context to mediation decisions and anonymization strength.

LLM-based text anonymization. Recent anonymization methods increasingly rewrite content rather than only tagging sensitive spans. Staab et al.’s adversarial anonymizer (Staab et al., 2025) iteratively rewrites text until a simulated attacker cannot infer attributes; AgentStealth (Shao et al., 2025) distils this adversarial loop with reinforcement learning; HaS (Chen et al., 2023b) sanitizes prompts before third-party LLM calls; CONFAIDE (Miresghallah et al., 2023) applies contextual-integrity prompts; and Rescriber (Zhou et al., 2025) uses a small LLM for on-device chat privacy. Related work also studies entity-tagging approaches (Mancera et al., 2025), clinical-text

anonymization (Pissarra et al., 2024), differential privacy for prompts (Duan et al., 2023), and private-attribute randomization (Frikha et al., 2025). These methods improve privacy protection beyond span detection, but adversarial rewriting loops often treat attacker confidence as direct evidence of residual leakage. This can trigger unnecessary rewriting when a guess is plausible under world knowledge but not actually supported by the original text.

Dual-system inference and uncertainty gating.

Routing only hard cases to an expensive model is a well-established efficiency strategy. Frugal-GPT (Chen et al., 2023a) cascades LLMs to reduce inference cost; BranchyNet (Teerapittayanon et al., 2016) adds early exits for neural networks; and fast/slow hybrid architectures draw on the broader dual-system view of decision-making (Kahneman, 2011). Evidential deep learning (Sensoy et al., 2018) provides a principled mechanism to estimate epistemic uncertainty and decide when to escalate. In privacy mediation, however, uncertainty gating cannot be only cost-driven: fast-path decisions must remain grounded in policy evidence, and uncertain cases must be escalated conservatively, as errors may expose sensitive information.

Positioning. GuardianAgent connects these lines of work by formulating live privacy mediation as policy-conditioned risk-adaptive anonymization. A structured risk controller links policy evidence and action context to both action-level decisions and anonymization strength, while verified adversarial feedback ensures that rewriting escalates only in response to residual leakage supported by the original text rather than hallucinated attacker confidence.

3 Method

3.1 Overview

A web action is represented as

$$\mathbf{a} = (d, \mathcal{D}_a, \mathcal{R}_a, \mathcal{P}_a, b_a, x_a) \in \mathcal{A},$$

where d is the destination domain, $\mathcal{D}_a$ the data categories (e.g., email, location, health, financial, or free-text content), $\mathcal{R}_a$ the recipients, $\mathcal{P}_a$ the declared purposes, $b_a \in \{\text{user-initiated, default, background}\}$ the action basis, and $x_a \in \Sigma^*$ the outgoing text. Here, *user-initiated* denotes an explicit user action such as submitting a form or pasting text, *default*

denotes routine processing needed for the requested service, and *background* denotes passive or script-triggered transmission without an immediate user action. Privacy policies are segmented into atomic statements $\mathcal{S}$, each annotated with structured fields such as data category, recipient, purpose, legal basis, retention, and user-rights disclosures. A retriever TOP_k selects the k most relevant policy statements as evidence $\mathcal{E} = \text{TOP}_k(\mathbf{a}, \mathcal{S})$. The matcher then emits an action decision and risk score, $(y, R) = M(\mathbf{a}, \mathcal{E})$, $\mathcal{Y} = \{\text{allow, transform, deny}\}$, and GuardianAgent implements

$$\Phi(\mathbf{a}, \mathcal{S}) = A(\mathbf{a}, y, R) \in \mathcal{Y} \times \Sigma^*. \quad (1)$$

If $y = \text{transform}$, the anonymizer A uses the transform-band position of R to select an initial rewrite level and then uses verified adversarial feedback to decide whether further escalation is needed. GuardianAgent separates two controls: AMRSF determines the starting strength from action context, while the verified guesser adjusts that strength only when residual leakage is surface-supported. The overall framework is introduced in Figure 1.

3.2 Efficient Policy Matching

The policy matcher estimates the likelihood that an outgoing action conflicts with the retrieved policy evidence. The likelihood L is then used as a learned input to the downstream risk controller R . To estimate L under the latency constraints of live web traffic, GuardianAgent uses a shared evidence retriever followed by a two-path matcher. Pre-segmented policies $s \in \mathcal{S}$ come with structured fields—data categories, actions, purposes, recipients, legal basis, rights, and retention—together with the original free-text snippet.

The evidence set $\mathcal{E} = \text{TOP}_k(\mathbf{a}, \mathcal{S})$ is retrieved using a hybrid score combining structured field overlap with lightweight lexical matching

$$\text{SCORE}(\mathbf{a}, s) = \alpha \tilde{S}(\mathbf{a}, s) + (1 - \alpha) \tilde{K}(\mathbf{a}, s), \quad (2)$$

where $\tilde{S}$ is the normalized structured score, $\tilde{K}$ is a normalized BM25Lite score, $\alpha = 0.6$, and $k = 8$.

The fast path applies an evidential classifier to the action–evidence pair and outputs both the risky-class mass and epistemic uncertainty u . When

$$u < \tau_u, \quad (3)$$

the risky-class mass is used as L . Otherwise, an LLM reasoning path estimates L from the same

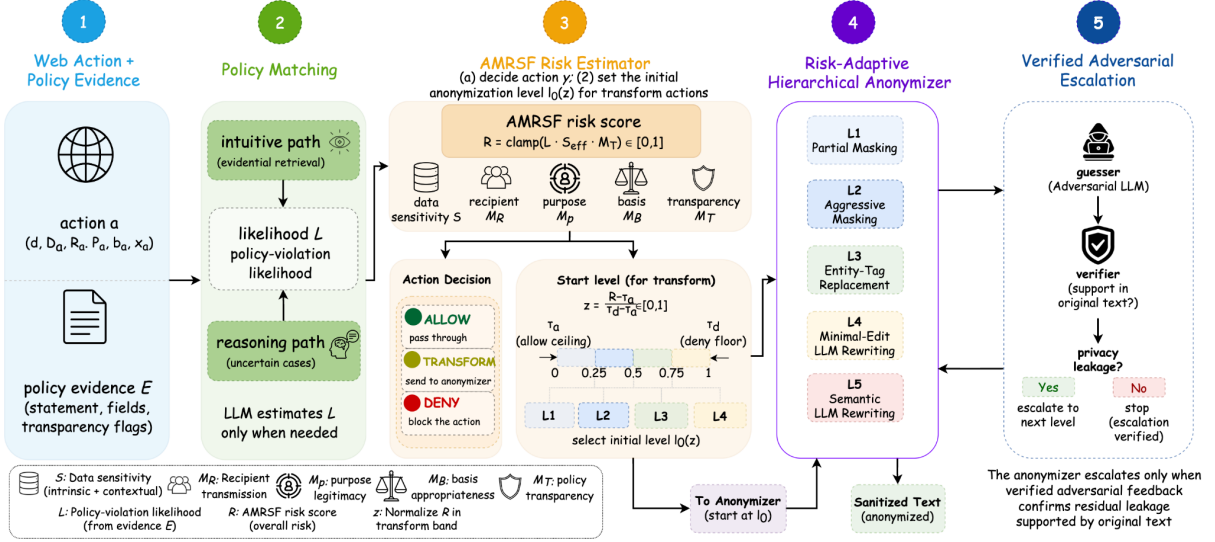


Figure 1: Framework of GuardianAgent. Policy matching estimates the violation likelihood L , and AMRSF converts it into an action-conditioned risk score R . The score has two effects: it determines the action decision $y \in \{\text{allow}, \text{transform}, \text{deny}\}$, and, for transformed actions, its normalized transform-band position z selects the initial anonymization level $\ell_0(z)$. The hierarchical anonymizer then escalates when leakage is detected.

action and evidence. Importantly, the LLM estimates only L : it does not directly determine the final risk score, action decision, or anonymization level. The encoder, retrieval weights, prompts, and online fast-path update are specified in Appendix A and Appendix B.

3.3 Policy-Conditioned Risk Estimator

The risk estimator converts the policy-violation likelihood L into an operational decision by accounting for the sensitivity and transmission context of the outgoing action. Given action $\mathbf{a}$, retrieved evidence $\mathcal{E}$, and the likelihood L estimated by the matcher, AMRSF computes the action-conditioned risk score

$$R(\mathbf{a}, \mathcal{E}) = \text{clamp}_{[0,1]}(L \cdot S_{\text{eff}} \cdot M_T), \quad (4)$$

where

$$S_{\text{eff}} = \min(1, S_d \cdot M_{tr} \cdot M_p \cdot M_b). \quad (5)$$

Here S_d is base data sensitivity; M_{tr} , M_p , and M_b adjust severity according to recipient transmission, purpose legitimacy, and action basis; and M_T accounts for policy transparency. Thus, unlike an LLM-assigned scalar risk judgment, R explicitly separates violation likelihood from the contextual severity of the outgoing action. These fixed factors are grounded in regulatory and empirical privacy-risk studies (McCallister, 2010; Regulation, 2016; Nissenbaum, 2009; Milne and Culnan, 2004; Ackerman et al., 1999; Bhatia and Breau, 2018). Full

numerical factors are reported in Tables 6–9 of Appendix C.

Let $t(\mathbf{a}) \in \{\text{crit}, \text{high}, \text{mod}, \text{low}\}$ denote the highest data-sensitivity tier in $\mathcal{D}_a$, with allow ceiling τ_a^t and deny floor τ_d^t . Then action decision

$$y = \begin{cases} \text{allow} & R < \tau_a^{t(\mathbf{a})}, \\ \text{transform} & \tau_a^{t(\mathbf{a})} \leq R < \tau_d^{t(\mathbf{a})}, \\ \text{deny} & R \geq \tau_d^{t(\mathbf{a})}. \end{cases} \quad (6)$$

Sensitive tiers use a wider transform region so that uncertain high-risk actions are rewritten rather than released unchanged. Threshold values are reported in Table 10 of Appendix C.

3.4 Risk-Adaptive Hierarchical Anonymizer

The hierarchical anonymizer controls how strongly transformed content is rewritten, using risk to select a starting strength and verified feedback to determine whether further escalation is needed. When $y = \text{transform}$, GuardianAgent rewrites the outgoing text x_a using a five-level risk-adaptive hierarchy: L1 partial masking, L2 aggressive masking, L3 entity-tag replacement, L4 minimal-edit LLM rewriting, and L5 semantic LLM rewriting. The initial level is determined by its normalized position within the transform band:

$$z(\mathbf{a}) = \text{clip}_{[0,1]} \left(\frac{R - \tau_a^{t(\mathbf{a})}}{\tau_d^{t(\mathbf{a})} - \tau_a^{t(\mathbf{a})}} \right), \quad (7)$$

and select the initial anonymization level by z :

Example Same text, increasing anonymization strength (L1 → L5)		Person / Family Organization Medical	Date / Time Transport Preference	Location / Place Work / Role Activity
Original	My son has an appointment at Riverside Children's Hospital this afternoon for an allergy test, so I'll leave work early after picking him up from Westbrook Primary School. We usually stop by the vegan cafe near Maple Station because he likes their mango smoothies. I'm also planning to meet my sister later for our weekly jazz class, and I still need to reply to my manager about tomorrow's night shift.	Contains medical, family, school, location, preference, hobby, and work clues		
L1 Partial Masking	My son has an appointment at Riverside Children's Hospital this afternoon for an allergy test, so I'll leave work early after picking him up from Westbrook Primary School. We usually stop by the vegan cafe near because he likes their mango smoothies. I'm also planning to meet my sister later for our [DATE] jazz class, and I still need to reply to my manager about [DATE]'s night shift.	Removes a few explicit spans, but most contextual clues remain		
L2 Aggressive Masking	My son has an appointment at this afternoon for an allergy test, so I'll leave work early after picking him up from . We usually stop by the vegan cafe near because he likes their mango smoothies. I'm also planning to meet my sister later for our [DATE] jazz class, and I still need to reply to my manager about [DATE]'s night shift.	Drop more specific spans, but still preserves the narrative structure		
L3 Entity-Tag Replacement	My son has an appointment at [ORG] this afternoon for an allergy test, so I'll leave work early after picking him up from [ORG]. We usually stop by the vegan cafe near [LOCATION] because he likes their mango smoothies. I'm also planning to meet my sister later for our [DATE] jazz class, and I still need to reply to my manager about [DATE]'s night shift.	Replaces sensitive entities with type-level tags while preserves the narrative structure		
L4 Minimal-Edit LLM Rewrite	My child has an appointment at a hospital this afternoon for a medical test, so I'll leave work early after picking them up from a school. We usually stop by a cafe near a train station because they like their smoothies. I'm also planning to meet a family member later for a class, and I still need to reply to my manager about a shift.	Generalizes sensitive spans into plausible but less identifying concepts		
L5 Semantic LLM Rewrite	I have an appointment this afternoon. I will leave work early to pick someone up from a school. We often stop by a cafe near a station. I have a meeting later for a class. I need to respond to a manager about a shift tomorrow.	Performs a broader semantic rewrite that keeps the main intent while removing many contextual cues		

Figure 2: Example of the five-level anonymization hierarchy on a contextually sensitive paragraph. From L1 to L5, more entities and contextual cues are anonymized, making the text less specific while preserving overall meaning.

$$\ell_0(z) = \begin{cases} 1 & z < 0.25, \\ 2 & 0.25 \leq z < 0.50, \\ 3 & 0.50 \leq z < 0.75, \\ 4 & z \geq 0.75. \end{cases} \quad (8)$$

Thus, transformed actions near the allow ceiling start with lighter rewriting, while those near the deny floor start from stronger levels. L1 partial masking, L2 aggressive masking, L3 entity-tag replacement, L4 minimal-edit LLM rewriting, and L5 semantic LLM rewriting. L1–L3 follow rule- and tag-based anonymization, L4 follows the minimal-edit paradigm, and L5 follows semantic iterative rewriting. Figure 2 illustrates how the hierarchy progressively removes specific entities and contextual cues while preserving the overall communicative intent. Lower levels preserve more of the original surface form and contextual detail, whereas higher levels are reserved for cases where weaker rewrites leave residual leakage.

3.5 Verified Adversarial Escalation

Prior adversarial anonymizers use an auxiliary attacker to identify residual personal information and guide further rewriting (Staab et al., 2025; Shao et al., 2025). However, an attacker may assign high confidence to a plausible guess that is unsupported by the original text, causing unnecessary escalation and utility loss. GuardianAgent therefore verifies attacker guesses before allowing them to trigger stronger rewriting. Given anonymized text x_ℓ , the attacker returns guesses with confidences $\mathcal{G}(x_\ell) = \{(g, c_g)\}$. The verifier compares each guess against the original text x_0 , which remains local to the anonymizer and is never exposed to the attacker. Let $\mathcal{W}(s)$ denote the set of content words in s with length greater than two. A guess is admitted only if it is explicitly contained in x_0 or shares a content word with it:

$$\mathcal{V}(g, x_0) = \mathbf{1}[g \subseteq x_0] \vee \mathbf{1}[\mathcal{W}(g) \cap \mathcal{W}(x_0) \neq \emptyset]. \quad (9)$$

The verified confidence is the maximum confidence among verified guesses:

$$\hat{c}(x_\ell, x_0) = \max\{c_g : (g, c_g) \in \mathcal{G}(x_\ell), \forall (g, x_0) = 1\}. \quad (10)$$

If no guess is verified, $\hat{c} = 0$. Starting from the initial level $\ell_0(z)$, GuardianAgent anonymizes at level ℓ , queries the attacker, and escalates to $\ell + 1$ only if $\hat{c} \geq \tau_g$. The loop terminates when $\hat{c} < \tau_g$, when $\ell = 5$, or after $T = 5$ rounds. Verification filters hallucinated or prior-driven guesses while preserving escalation signals for leakage that is supported by the original text. Equation 9 is explicitly lexical: it does not recognize a correct contextual inference that is surface-disjoint from x_0 . We therefore treat it as an anti-hallucination and escalation-cost control rather than as the system’s contextual-inference defence; higher-risk flows instead start at L4 through R and receive semantic rewriting without waiting for the verifier. The pseudocode, lexical verification guarantee, and the expected-cost analysis are in Appendix D.

4 Evaluation

The evaluation addresses four questions: (Q1) Does GuardianAgent improve the privacy–utility trade-off? (Q2) Does action-conditioned risk control both mediation decisions and anonymization strength? (Q3) Does verified hierarchical anonymization outperform fixed rewriting strategies? (Q4) Are the risk controller and intuitive+reasoning path calibrated and efficient? Unless otherwise stated, all main results use the final GuardianAgent variant with the full L1–L5 hierarchy. Calibration results, additional confidence intervals, rewrite strategy comparison, seed variance, and robustness checks are reported in Appendix E, F and G.

4.1 Setup

Benchmarks. Our evaluation combines anonymization, policy-matching, and risk-calibration data. For anonymization, we use three benchmarks **TAB** (Pilán et al., 2022), which contain legal texts from ECHR judgements; **SynthPAI** (Yukhymenko et al., 2024), which contains Reddit-style posts with personal attributes; and **PII-Masking-300k** (A4Privacy, 2024), which contains XML, JSON, tabular, and free-form synthetic documents annotated with explicit PII categories. For policy matching, we use **OPP-115** (Wilson et al., 2016). For AMRSF calibration, we use an expert-rated action-scenario set and the public corpus from Staab.

Backbone and baselines. The main LLM backbone is Llama-3.2-3B-Instruct (Grattafiori et al., 2024); System 2, L4 rewriting, and the adversarial guesser use the same backbone to ensure a controlled comparison across components. We compare against baselines spanning adversarial rewriting, prompt protection, contextual-integrity prompting, on-device rewriting, differential privacy, private-attribute randomization, clinical anonymization, and tag-based anonymization: Staab (Staab et al., 2025), AgentStealth (Shao et al., 2025), HaS (Chen et al., 2023b), CONFAIDE (Mireshghallah et al., 2023), Rescriber (Zhou et al., 2025), DP-Prompt (Duan et al., 2023), IncogniText (Frikha et al., 2025), Pissarra et al. (Pissarra et al., 2024), and PBa-LLM (Mancera et al., 2025).

Metrics. *Privacy* is one minus the fraction of ground-truth sensitive spans retained in the output. *Utility* is token-Jaccard on TAB and PII-Masking-

Method	TAB (legal, $n = 811$)				SynthPAI (Reddit, $n = 200$)				PII-Masking-300k ($n = 500$)			
	Priv↑	Util↑	Guess↓	Lat.	Priv↑	Util↑	Guess↓	Lat.	Priv↑	Util↑	Guess↓	Lat.
<i>Iterative / multi-pass LLM rewriters</i>												
A. Ours (full, L1–L5)	.945 _{.010}	.696 _{.008}	.283 _{.019}	5052 ₁₈₁	.944 _{.007}	.888 _{.006}	.295 _{.009}	9949 ₄₆₇	.901 _{.011}	.488 _{.008}	.213 _{.010}	7453 ₂₉₉
B. Staab’25	.695 _{.013}	.564 _{.017}	.335 _{.019}	25152 ₂₃₉	.975 _{.005}	.257 _{.018}	.098 _{.014}	45418 ₅₆₆₃	.850 _{.010}	.384 _{.021}	.275 _{.020}	24975 ₉₆
C. AgentStealth’25	.689 _{.012}	.789 _{.007}	.579 _{.044}	16928 ₅₆₃	.864 _{.009}	.700 _{.005}	.260 _{.009}	15485 ₄₆₇	.680 _{.008}	.634 _{.015}	.508 _{.015}	19414 ₃₉₄
D. HaS’23	.926 _{.005}	.669 _{.009}	.336 _{.032}	5328 ₅₉₄	.811 _{.026}	.841 _{.005}	.477 _{.020}	12892 ₇₅₅	.809 _{.010}	.586 _{.011}	.322 _{.015}	8384 ₂₈₅
E. IncogniText’25	.178 _{.023}	.765 _{.011}	.538 _{.025}	8341 ₃₂₅	.553 _{.019}	.855 _{.013}	.281 _{.023}	10092 ₄₁₃	.315 _{.011}	.701 _{.003}	.624 _{.018}	9819 ₂₀₂
<i>Single-shot LLM rewriters</i>												
F. CONFAIDE’24	.913 _{.005}	.638 _{.008}	.307 _{.027}	679 ₉₇	.842 _{.013}	.810 _{.010}	.254 _{.015}	2955 ₃₀₉	.776 _{.009}	.533 _{.004}	.331 _{.012}	2449 ₁₈₂
G. Pissarra’24	.903 _{.010}	.701 _{.005}	.257 _{.030}	789 ₁₂₇	.815 _{.002}	.873 _{.003}	.395 _{.006}	2943 ₂₉₃	.864 _{.010}	.548 _{.008}	.256 _{.013}	2846 ₁₅₇
H. Rescriber’25	.719 _{.022}	.605 _{.031}	.393 _{.022}	3774 ₃₉₈	.819 _{.012}	.432 _{.020}	.182 _{.012}	3241 ₃₄₅	.892 _{.009}	.282 _{.008}	.178 _{.013}	2752 ₇₆
I. DP-Prompt’23	.375 _{.029}	.478 _{.015}	.662 _{.019}	837 ₁₄₆	.486 _{.019}	.539 _{.032}	.319 _{.017}	3371 ₄₂₉	.524 _{.025}	.527 _{.008}	.505 _{.012}	3611 ₉₇
<i>Tag-based baselines</i>												
J. Presidio	.703 _{.015}	.783 _{.008}	.216 _{.023}	14 ₁	.659 _{.005}	.944 _{.001}	.375 _{.015}	28 ₁	.511 _{.019}	.750 _{.011}	.489 _{.022}	24 ₁
K. PBa-LLM’25	.455 _{.033}	.809 _{.009}	.465 _{.019}	81 ₁₃	.635 _{.002}	.956 _{.000}	.396 _{.008}	170 ₃₃	.184 _{.013}	.906 _{.005}	.674 _{.013}	91 ₁

Table 1: Main comparison across TAB, SynthPAI, and PII-Masking-300k. All three corpora report 4-seed mean_{std} (42–45). Latency is mean ms per sample. **Best** and second-best are marked within each method block and dataset. [†]Staab’s high SynthPAI privacy is inflated by empty or placeholder outputs, which also explains its low utility.

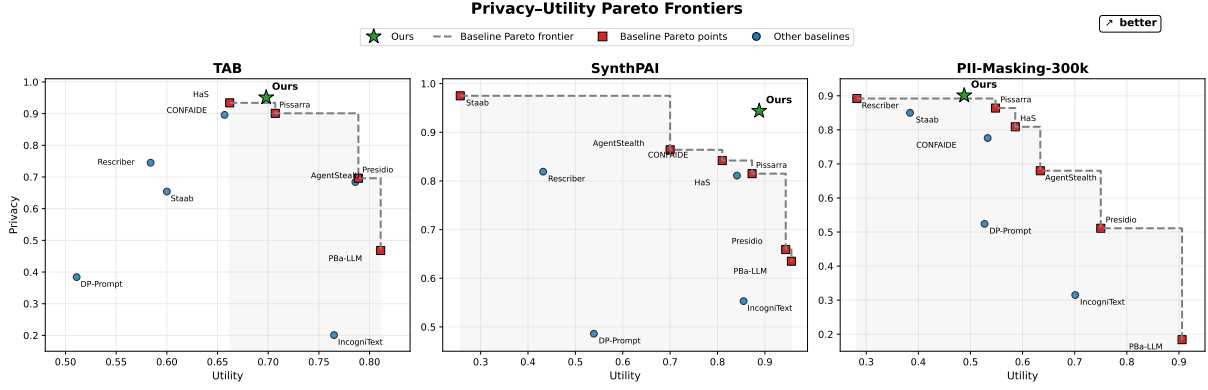


Figure 3: Privacy–utility Pareto frontiers across the three benchmarks. Baseline methods define each frontier, with GuardianAgent plotted separately. GuardianAgent lies above or beyond the baseline frontier on all three datasets, indicating a stronger cross-domain privacy–utility trade-off.

300k, and BERTScore F1 on SynthPAI. *Guess* is the maximum verified adversarial-guesser confidence on the output, and *latency* is the wall-clock time per sample.

4.2 Main Comparison

Table 1 compares GuardianAgent with published baselines across TAB, SynthPAI, and PII-Masking-300k. GuardianAgent delivers the most consistent, privacy–utility trade-off across domains. Figure 3 reinforces this result: GuardianAgent lies above or beyond the baseline privacy–utility frontier on all three datasets, and is the only method above 0.90 privacy on all domains and has the smallest cross-domain spread. This consistency suggests that verified, risk-adaptive escalation reduces domain-specific failure modes, especially when moving across legal prose, Reddit-style personal attributes, and structured PII records.

Table 2 evaluates the decision side of the controller by scoring the same 200 PII-Masking texts under four action contexts with $L = 0.7$ fixed. As recipient, purpose, action basis, and policy transparency become riskier, mean risk increases and decision distribution shifts from allow through transform to deny. This shows that AMRSF con-

ditions mediation decisions on action context rather than on text content alone.

4.3 Transform-Band Risk-to-Level Control

To isolate anonymization strength, the same 200 PII-Masking texts are evaluated under four contexts constrained to the transform band. For each action, z denotes the normalized position of R between the allow ceiling and deny floor (Eq. 7), and $\ell_0(z)$ is selected by quartile (Eq. 8).

Table 3 shows that anonymization strength follows the action-conditioned band position. As z increases, the dominant starting level shifts from L1 to L4 and $\ell_0(z)$ rises monotonically. Verified adversarial escalation preserves this ordering, with the mean final level rising from 2.37 to 4.42. Thus, among actions selected for transformation, GuardianAgent does not use a fixed rewriting policy: AMRSF sets the starting strength from the action context, and verified feedback escalates only when residual leakage remains.

4.4 Ablation

Table 5 decomposes the contributions of the main components. NER alone is strong on TAB, where many sensitive spans are explicit legal entities, but

Context	R	Allow	Trans.	Deny	$\bar{\ell}_{cf}$
C_{low} (first-party / functionality / paste / full)	.119 _{.017}	100.0	0.0	0.0	2.02 _{1.31}
C_{med} (provider / analytics / default / partial)	.413 _{.057}	0.5	98.0	1.5	2.10 _{1.29}
C_{high} (advertising / profiling / background / vague)	.909 _{.016}	0.0	0.5	99.5	4.33 _{0.50}
C_{ext} (data broker / advertising / background / missing)	.909 _{.011}	0.0	0.0	100.0	4.34 _{0.49}

Table 2: Action-conditioned decision control on 200 PII-Masking samples with the same texts and fixed $L = 0.7$. Riskier action contexts shift decisions from allow to transform and deny. $\bar{\ell}_{cf}$ is a counterfactual final anonymization level computed by routing all samples through the anonymizer; at deployment time, allow samples are released unchanged, while deny samples are blocked.

Context	Recipient	Purpose	Basis	R	z	Trans. %	$\ell_0(z)$	L1	L2	L3	L4	ℓ_{final}
T_{mild}	first-party	Personalization	default	.340 _{.050}	.252 _{.179}	99.0	1.48	65	22	13	0	2.37
T_{moderate}	service-provider	Personalization	default	.457 _{.066}	.538 _{.208}	89.0	2.52	1	64	16	18	2.93
T_{strong}	service-provider	Marketing	default	.557 _{.082}	.744 _{.170}	81.5	3.33	1	0	64	35	3.64
$T_{\text{near-deny}}$	advertising	Marketing	default	.704 _{.053}	.980 _{.080}	13.0	3.98	0	0	0	99	4.42

Table 3: Transform-band risk-to-level control on 200 PII-Masking samples with fixed $L = 0.7$. As normalized band position z increases, the dominant starting level shifts from L1 to L4, the mean initial level $\ell_0(z)$ rises monotonically, and the final level after verified escalation also increases. Trans. % is the fraction of samples remaining within the transform region.

Cap	TAB			SynthPAI			PII-Mask		
	Priv	Util	Guess	Priv	Util	Guess	Priv	Util	Guess
$\ell_{\text{max}}=3$	.924	.709	.356	.913	.897	.329	.869	.533	.247
$\ell_{\text{max}}=4$	.933	.706	.325	.912	.905	.316	.849	.539	.283
$\ell_{\text{max}}=5$	.945 _{.011}	.696 _{.009}	.296	.944 _{.007}	.888 _{.006}	.299	.901 _{.011}	.488 _{.008}	.213 _{.010}

Table 4: Level-progression ablation with maximum level $\ell_{\text{max}} \in \{3, 4, 5\}$. L3 uses only rule-based masking/tagging, L4 adds minimal-edit rewriting, and L5 adds the semantic safety-net.

Variant	Priv $\uparrow$	Util $\uparrow$	ms
<i>SynthPAI</i>			
A. NER only	.483 _{.003}	.944 _{.001}	23
B. + guesser	.569 _{.007}	.938 _{.001}	729
C. + LLM rewrite	.910 _{.011}	.899 _{.003}	2518
D. + L1-L5 adaptive	.944 _{.007}	.888 _{.006}	9949
<i>TAB</i>			
A. NER only	.848 _{.022}	.735 _{.009}	11
B. + guesser	.895 _{.013}	.718 _{.007}	2731
C. + LLM rewrite	.915 _{.018}	.725 _{.010}	831
D. + L1-L5 adaptive	.945 _{.011}	.696 _{.009}	5052
<i>PII-Masking-300k</i>			
A. NER only	.389 _{.014}	.813 _{.009}	14
B. + LLM rewrite	.853 _{.007}	.546 _{.008}	4084
C. + L1-L5 adaptive	.901 _{.011}	.488 _{.008}	7453

Table 5: Component ablation on SynthPAI, TAB, and PII-Masking-300k (mean_{std} across 4 seeds; $n = 200$ for SynthPAI/TAB and $n = 500$ for PII-Masking). Variants are cumulative: NER only, plus adversarial feedback, plus single-level LLM rewriting, and the full verified L1-L5 adaptive stack.

it is less effective on SynthPAI, where personal attributes are more implicit and contextual. Adding adversarial feedback and LLM rewriting improves privacy, while the full L1-L5 adaptive stack achieves the best privacy across both datasets at a moderate utility cost. Additional comparisons between semantic-only and minimal-edit-only rewriting are reported in Table 13 (Appendix G). Table 4 isolates the role of the anonymization hierarchy. L1-L3 remove many explicit spans, especially in legal prose and structured records. L4 adds high-utility minimal edits, while L5 is most useful when weaker rewrites leave contextual leakage. The results on PII-Masking-300k shows why a fixed max-

imum level can be brittle: minimal edits may preserve too much structure, whereas the semantic safety-net recovers privacy in harder cases. Qualitative side-by-side rewrites are in Appendix H.

4.5 Calibration, Efficiency, and Robustness

AMRSF is evaluated on an expert-rated scenario set and a larger Staab-derived calibration corpus. The score aligns well with annotated risk in both settings, while the larger corpus provides a more stable estimate because the expert-rated set is small. The evidential fast path is evaluated on OPP-115 policy-action pairs and escalates only a small fraction of events to the LLM, supporting its role as an efficiency mechanism for live traffic mediation. Full calibration statistics, confidence intervals, and System 1 metrics are reported in Appendix E.

Robustness checks support the two adaptive mechanisms. Disabling verification leads to more level upgrades without improving privacy, while reducing utility and increasing latency, confirming that raw attacker confidence often reflects unsupported guesses. By switching backbone from Llama-3.2-3B to Qwen3-8B, GuardianAgent remains the strongest privacy-utility method on test benchmarks. Additional paired-bootstrap tests, seed variance, attacker-replacement experiments, parsing-failure analysis, backbone-portability details, and online fast-path self-improvement results are in Appendix G.

5 Conclusion

This paper presented GuardianAgent, a policy-conditioned framework for risk-adaptive

anonymization of live web traffic. Rather than relying on an LLM to directly assign privacy risk, GuardianAgent uses AMRSF to combine policy evidence and action context, determining both the allow/transform/deny decision and the initial anonymization level. A verified adversarial guesser then escalates rewriting only when attacker guesses are supported by the original text. Across legal text, Reddit-style personal attributes, and multi-format PII records, GuardianAgent improved the privacy–utility trade-off over published baselines, remained robust under a backbone switch, and adapted its rewriting strength across different action contexts.

Limitations

The evaluation is limited to English and uses public or synthetic corpora; real browser traces, clinical and dialogue data, and multilingual settings require further study. AMRSF should also be viewed as a mediation controller rather than a legal compliance engine. Deployments should therefore prioritize local processing, disclose any cloud escalation, allow user override, and communicate that anonymization reduces but does not eliminate the re-identification risk.

Ethical considerations

This work collects no new personal data and involves no human subjects. Experiments use publicly released or synthetic benchmark datasets and an open-weight backbone or publicly accessible model APIs; no private browser traces are used. GuardianAgent is intended to reduce unintended disclosure, not to guarantee anonymity or legal compliance. Any deployment should disclose rewriting and cloud escalation, preserve user control over mediation decisions, and communicate residual re-identification and utility-loss risks.

Acknowledgments

The authors thank the UNSW-UTS Trustworthy Digital Society for supporting this research. Also, this research used the NVIDIA H100 cluster provided by SHARON AI.

References

Mark S Ackerman, Lorrie Faith Cranor, and Joseph Reagle. 1999. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings*

of the 1st ACM Conference on Electronic Commerce, pages 1–8.

Wasi Ahmad, Jianfeng Chi, Tu Le, Thomas Norton, Yuan Tian, and Kai-Wei Chang. 2021. Intent classification and slot filling for privacy policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4402–4417.

Ai4Privacy. 2024. [pii-masking-300k \(revision 86db63b\)](#).

Jaspreet Bhatia and Travis D Breaux. 2018. Empirical measurement of perceived privacy risk. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(6):1–47.

Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-jun Li, and Yaxing Yao. 2025. Clear: Towards contextual llm-empowered privacy policy analysis and risk generation for large language model applications. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 277–297.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023a. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.

Yu Chen, Tingxin Li, Huiming Liu, and Yang Yu. 2023b. Hide and seek (has): A lightweight framework for prompt privacy protection. *arXiv preprint arXiv:2309.03057*.

R Jason Cronk and Stuart S Shapiro. 2021. Quantitative privacy risk analysis. In *2021 IEEE European symposium on security and privacy workshops (EuroS&pW)*, pages 340–350. IEEE.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. 2023. Flocks of stochastic parrots: Differentially private prompt learning for large language models. *Advances in Neural Information Processing Systems*, 36:76852–76871.

Vincent Freiburger, Arthur Fleig, and Erik Buchmann. 2026. Helping johnny make sense of privacy policies with llms. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2025. Incognitext: Privacy-enhancing conditional text anonymization via llm-based private attribute randomization. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2490–2501.

- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548.
- Daniel Kahneman. 2011. Fast and slow thinking. *Allen Lane and Penguin Books, New York*, 2.
- Gonzalo Mancera, Aythami Morales, Julian Fierrez, Ruben Tolosana, Alejandro Peña, Miguel Lopez-Duran, Francisco Jurado, and Alvaro Ortigosa. 2025. Pba-llm: Privacy-and bias-aware nlp using named-entity recognition (ner). In *International Conference on Document Analysis and Recognition*, pages 3–20. Springer.
- Erika McCallister. 2010. *Guide to protecting the confidentiality of personally identifiable information*. Diane Publishing.
- George R Milne and Mary J Culnan. 2004. Strategies for reducing online privacy risks: Why consumers read (or don’t read) online privacy notices. *Journal of interactive marketing*, 18(3):15–29.
- Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.
- Ivoline C Ngong, Swanand Ravindra Kadhe, Hao Wang, Keerthiram Murugesan, Justin D Weisz, Amit Dhurandhar, and Karthikeyan Natesan Ramamurthy. 2025. Protecting users from themselves: Safeguarding contextual privacy in interactions with conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26196–26220.
- Helen Nissenbaum. 2009. Privacy in context: Technology, policy, and the integrity of social life. In *Privacy in context*. Stanford University Press.
- Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.
- David Pissarra, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André Carreiro, and Vitor Rolla. 2024. Unlocking the potential of large language models for clinical text anonymization: A comparative study. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 74–84.
- Protection Regulation. 2016. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679(2016):10–3.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.
- Chenyang Shao, Tianxing Li, Chenhao Pu, Fengli Xu, and Yong Li. 2025. Agentstealth: Reinforcing large language model for anonymizing user-generated text. *arXiv preprint arXiv:2506.22508*.
- Mukund Srinath, Soundarya Nurani Sundareswara, C Lee Giles, and Shomir Wilson. 2021a. Privaseer: A privacy policy search engine. In *International conference on web engineering*, pages 286–301. Springer.
- Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021b. Privacy at scale: Introducing the privaseer corpus of web privacy policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6829–6839.
- Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2025. Language models are advanced anonymizers. In *The Thirteenth International Conference on Learning Representations*.
- Bolun Sun, Yifan Zhou, and Haiyun Jiang. 2025. Empowering users in digital privacy management through interactive llm-based agents. In *International Conference on Learning Representations*, volume 2025, pages 76740–76760.
- Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xi-ang Li, Tianming Liu, and Lei Fan. 2023. [Policygpt: Automated analysis of privacy policies with large language models](#). *Preprint*, arXiv:2309.10238.
- Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2464–2469. IEEE.
- Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard

Hovy, Joel Reidenberg, and Norman Sadeh. 2016. [The creation and analysis of a website privacy policy corpus](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, Berlin, Germany. Association for Computational Linguistics.

Qinge Xie, Karthik Ramakrishnan, and Frank Li. 2025. Evaluating privacy policies under modern privacy laws at scale: An {LLM-Based} automated approach. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 5797–5816.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. 2024. A synthetic dataset for personal attribute inference. *Advances in Neural Information Processing Systems*, 37:120735–120779.

Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025. Rescriber: Smaller-llm-powered user-led data minimization for llm-based chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

A System 1 Architecture and Retrieval Hyperparameters

System 1 encoder and EDL classifier (referenced from §3.2). Behaviour and policy encoders $\phi_b, \phi_p : \mathcal{A} \cup \mathcal{S} \rightarrow \mathbb{R}^d$ are either a $d = 42$ multi-hot variant over OPP-115 categories or a $d = 384$ frozen sentence-transformer (Reimers and Gurevych, 2019). Writing $b = \phi_b(\mathbf{a})$ and $p = \phi_p(\mathbf{s})$, the interaction feature map is

$$\psi(\mathbf{a}, \mathbf{s}) = [b; p; b \odot p; |b - p|] \in \mathbb{R}^{4d}. \quad (11)$$

A two-layer MLP with ReLU and dropout, $f_\theta : \mathbb{R}^{4d} \rightarrow \mathbb{R}^K$ with $K = 2$, yields logits that we convert to Dirichlet parameters following EDL (Sensoy et al., 2018): $\alpha = \text{softplus}(f_\theta(\psi)) + 1$, $S = \sum_k \alpha_k$. Class probabilities are $p_k = \alpha_k / S$ and epistemic uncertainty is $u = K / S$. The likelihood used in Eq. 4 is $L = \alpha_2 / S$.

Retrieval scoring (referenced from §3.2). The structured score over fields $\mathcal{F} = \{\text{data}, \text{act}, \text{purp}, \text{recip}\}$ is

$$S(\mathbf{a}, s) = \sum_{f \in \mathcal{F}} w_f \rho_f + \text{FLAGS}(s), \quad (12)$$

with weights $(0.35, 0.35, 0.20, 0.10)$, $\rho_f = \min(1, |\mathbf{a}_f \cap s_f| / \max(1, |\mathbf{a}_f|))$, and $\text{FLAGS}(s)$ adding small bonuses for disclosed user rights, trans-EEA transfer, and retention. BM25Lite uses $k_1 = 1.5$, $b = 0.75$ over the tokenized snippet, with on-the-fly IDF (no global inverted index per-domain).

B Prompts

System 2 decider (system prompt).

Given an outgoing action and retrieved policy evidence, estimate the likelihood that the action conflicts with the policy. Return STRICT JSON: {violation_likelihood: [0..1], rationale, evidence_ids}. Do not decide whether the action should be allowed, transformed, or denied.

Adversarial guesser (system prompt).

You are an adversarial privacy auditor. Given anonymized text, try to reconstruct or guess the redacted information. For each guess, provide: field, guess, confidence (0.0–1.0), reasoning. Return JSON: {"guesses": [...], "max_confidence": 0.0–1.0}.

L4 minimal-edit rewriter.

Rewrite the following sentence so that the listed sensitive fields cannot be recovered, while preserving the communicative intent. Do not insert [TAGS]; produce natural text. Sensitive fields: {fields}. Input: "{text}".

L5 semantic rewriter.

Rewrite the text to preserve the communicative intent while removing identifying details and contextual cues. Full paraphrasing is allowed.

C Full AMRSF Factor Tables

Table 6 gives the full per-category data weights used to compute S_d in Eq. 5. Table 7 gives the per-recipient transmission risk M_{tr} and Table 8 the per-purpose legitimacy M_p . Table 9 lists the six-signal credits used by the transparency multiplier M_T .

Average final level (ablation). On TAB, the verified loop (D) reduces the mean final anonymization level from 3.48 (C) to 2.85 (D). On SynthPAI, it rises slightly from 2.79 to 3.48 as the verifier correctly triggers upgrades on borderline inputs.

D Formal Analysis

Proposition 1 (No missed lexically-supported guesses). *Assume a guess g identifies an attribute of x_0 only if it either appears as a substring of x_0 or shares at least one content word (> 2 characters) with x_0 . Then $\mathcal{C}(x_0) \subseteq \mathcal{V}(x_0)$, and consequently the verified confidence $\hat{c}$ in Eq. 10 satisfies $\hat{c} \geq \max_{g \in \mathcal{C}(x_0)} c_g$.*

Category	Wt.	Source
Credentials	1.00	NIST 800-122
Financial	0.95	Milne'04
Health, Biometric, Genetic	0.95	GDPR Art. 9
SSN	0.90	NIST 800-122
Location (precise)	0.85	Milne'04
Location (unspec.)	0.75	default
Phone	0.70	Ackerman'99
Content	0.65	Bhatia'18
PostalAddress	0.65	Ackerman'99
Contact	0.60	default
Email	0.55	Ackerman'99
SearchHistory	0.50	Milne'04
BrowsingHistory	0.45	Milne'04
IPAddress, DeviceID	0.40	NIST
Location (coarse)	0.35	Milne'04
AppUsage	0.25	low
Cookies	0.20	low
Language	0.10	low

Table 6: Full data-sensitivity weights $DW(c)$. Some rows group categories assigned the same weight.

Recipient	M_{tr}
first-party only	0.20
service provider (Art. 28)	0.40
analytics / government	0.50
social-media platform	0.60
unknown third party	0.70
advertising network	0.80
data broker	0.90

Table 7: Transmission risk s_{tr} by recipient type.

Proof. If $g \in \mathcal{C}(x_0)$ matches by substring, the first disjunct of Eq. 9 fires; if by shared content word, the second disjunct fires. In either case $V(g, x_0) = 1$, so $g \in \mathcal{V}$. Taking max over the superset bounds $\max_{g \in \mathcal{C}} c_g$ from above. $\square$

Proposition 2 (Bounded expected cost of the gate). *Let C_1, C_2 be the per-event costs of System 1 and System 2 and $p_e = \Pr[u \geq \tau_u]$ the escalation probability. Then the expected matcher cost is $\mathbb{E}[C] = C_1 + p_e C_2$. With the values measured on OPP-115 ($p_e = 0.035$) and $C_2/C_1 \approx 10^4$, $\mathbb{E}[C]/C_2 \approx p_e$, a 28.6 $\times$ reduction over an always-LLM pipeline.*

Proof. System 1 runs unconditionally; Eq. 3 makes System 2 run exactly when $u \geq \tau_u$, a Bernoulli event of rate p_e . Linearity of expectation gives the result. $\square$

Proposition 3 (Loop termination). *Algorithm 1 halts in at most T rounds and invokes at most one anonymizer call plus one guesser call per round, so its worst-case LLM cost is $T \cdot (C_{anon} + C_{guess})$ —a constant independent of x_0 .*

Proof. Each iteration either breaks (on $\hat{c} < \tau_g$

Purpose	M_p
functionality, security, auth.	0.15
personalization, research	0.35
analytics	0.45
marketing	0.65
profiling (Art. 22)	0.75
advertising	0.80
undisclosed	0.65

Table 8: Purpose legitimacy s_p .

Signal	Credit
substantial text (≥ 50 chars)	−0.10
covers the action’s data category	−0.15
covers the action’s declared purpose	−0.10
states a legal basis (GDPR Art. 6)	−0.08
discloses user rights	−0.05
specifies retention	−0.02
$M_T = \max(0.8, 1.3 + \sum \text{credits})$	

Table 9: Transparency multiplier M_T signal credits.

or $\ell = \ell_{\max}$) or increments ℓ . Since $\ell \leq \ell_{\max} = 5$ and the explicit cap $T = 5$ is enforced, termination follows. $\square$

E Calibration Details

Table 11 compares the full AMRSF controller against three single-factor ablations, a NIST Likelihood $\times$ Impact baseline, and a learned binary classifier on the expert-rated scenarios and the Staab public corpus. Table 12 reports the System 1 evidential fast-path classifier on the OPP-115 held-out test pairs.

F Rewrite-strategy comparison

Table 13 contrasts two single-strategy variants of D (semantic-only rewriting and minimal-edit-only rewriting) with the full adaptive L1–L5 stack used by GuardianAgent, isolating the effect of mixing rewrite strategies under the verified-guesser loop.

G Robustness Detail Tables

Table 14 reports an A/B test that turns the verifier on or off in the adaptive loop, isolating its contribution to the privacy/utility trade-off and to escalation cost.

Table 15 reports four-seed variance for the five strongest baselines plus GuardianAgent, so the per-method spread can be read off directly.

Table 16 swaps the LLM backbone between Llama-3.2-3B and Llama-3.1-8B to check that

Data tier	Allow ceiling τ_a^t	Deny floor τ_d^t
Critical	0.10	0.55
High	0.25	0.65
Moderate	0.35	0.75
Low	0.45	0.80

Table 10: Tier-specific thresholds for the action decision in Eq. 6. Tiers are assigned to data categories as: *Critical* (Credentials, Health, Biometric, Genetic, Financial, SSN), *High* (precise Location, Phone, Content), *Moderate* (Email, Contact, SearchHistory, BrowsingHistory), *Low* (Cookies, AppUsage, DeviceID, Language). The transform region width $\tau_d^t - \tau_a^t$ is widest for sensitive tiers, biasing borderline actions on sensitive data toward rewriting rather than passing through unchanged.

Algorithm 1: Verified adaptive anonymization (VERIFIEDADAPT)

Require: original text x_0 ; risk R ; data tier $t(\mathbf{a})$; thresholds (τ_a^t, τ_d^t) ; guesser $\mathcal{G}$; verifier $\mathbf{V}$; $\tau_g=0.6$; $\ell_{\max}=5$; $T=5$

- 1: $z \leftarrow \text{clip}_{[0,1]} \left(\frac{R - \tau_a^{t(\mathbf{a})}}{\tau_d^{t(\mathbf{a})} - \tau_a^{t(\mathbf{a})}} \right)$
- 2: $\ell \leftarrow \ell_0(z)$
- 3: **for** $t = 1, \dots, T$ **do**
- 4: $x_\ell \leftarrow \text{Anonymize}(x_0, \ell)$
- 5: **if** $\ell = \ell_{\max}$ **then break**
- 6: **end if**
- 7: $\mathcal{G}_\ell \leftarrow \mathcal{G}(x_\ell)$
- 8: $\mathcal{G}_V \leftarrow \{(g, c_g) \in \mathcal{G}_\ell : \mathbf{V}(g, x_0) = 1\}$
- 9: $\hat{c} \leftarrow 0$ **if** $\mathcal{G}_V = \emptyset$, **else** $\max_{(g, c_g) \in \mathcal{G}_V} c_g$
- 10: **if** $\hat{c} \geq \tau_g$ **then** $\ell \leftarrow \ell + 1$
- 11: **else break**
- 12: **end if**
- 13: **end for**
- 14: **return** x_ℓ, ℓ

Figure 4: Verified adaptive anonymization. The initial level is selected from the transform-band position z , and the verifier admits only guesses supported by the original text.

GuardianAgent’s ranking is not an artifact of one model size.

Table 17 buckets GuardianAgent’s outputs into four error classes to characterize where the system fails when it fails. After Holm–Bonferroni correction at family-wise $\alpha = 0.05$, 30/40 pairwise tests on TAB and 28/40 on SynthPAI remain significant.

Table 18 repeats the main comparison with Qwen3-8B as the LLM backbone for every method, providing a cross-model-family check that Ours’s privacy ranking is not Llama-specific.

Table 19 tracks the System 1 fast path as it is online-bootstrapped from System 2 teacher labels on an OPP-115 event stream, showing how quickly the slow-path escalation rate drops as fast-path accuracy climbs.

Method	ρ	r	MAE	Macro-F1
<i>Expert-rated scenarios (n = 500)</i>				
AMRSF (full)	0.982	0.947	0.119	0.812
AMRSF – M_b	0.965	0.940	0.136	0.683
AMRSF – M_T	0.985	0.956	0.134	0.738
AMRSF – purpose	0.976	0.947	0.135	0.738
NIST L×I	0.964	0.949	0.108	0.756
Binary classifier	0.729	0.683	0.338	0.435
<i>Staab public corpus (n = 525)</i>				
AMRSF (full)	0.914	0.866	0.268	–

Table 11: Risk-score calibration on the expert-rated scenario set and Staab public corpus. The larger Staab corpus provides a more stable estimate because the expert-rated set is small.

Metric	Value
Test pairs (OPP-115)	6,392
Accuracy	0.928
Precision	1.000
Recall	0.857
F_1	0.923
Expected Calibration Error	0.048
Brier score	0.058
Escalation rate ($u \geq 0.25$)	3.47%
p95 latency (fast path)	0.25 ms

Table 12: System 1 evidential classifier on OPP-115.

H Qualitative Case Studies

Two verbatim examples illustrate flagship-D behaviour relative to baselines. On the TAB sentence “On 13 September 1988 the Erzincan Martial Law Court acquitted the applicants of the charges against them”: GuardianAgent (L4) emits “On a date in 1988 the court acquitted the applicants of the charges against them”, swapping the specific date and court identity while keeping the legal speech-act intact. Staab rewrites to two sentences with redundant information (“On a certain month, a martial law court acquitted... In a certain jurisdiction, a court acquitted...”), duplicating the act; HaS returns “On [DATE] [YEAR] the [PLACE] Martial Law Court acquitted the applicants...”, leaving the court category intact with a tag-heavy output that hurts fluency. On a SynthPAI post where a software engineer describes their Masters in Computer Science and their salary in Zurich, GuardianAgent rewrites to generic remarks about graduate studies and work while Staab preserves the Zurich reference by hallucinating “a European city”—a regional clue that still narrows the author.

D variant	TAB		SynthPAI	
	Priv	Util	Priv	Util
D (semantic only)	.974	.628	.942	.847
D (min-edit only)	.866	.740	.892	.892
D (L1–L5 adaptive)	.945 _{.011}	.696 _{.009}	.944 _{.007}	.888 _{.006}

Table 13: Rewrite-strategy comparison for variant D. The adaptive L1–L5 row reports mean_{std} across 4 seeds at $n = 200$; the two pilot rows are single-seed.

Metric	V ON	V OFF	Δ
Final level	3.52	3.96	−0.44
Upgrade rate	0.52	0.80	−0.28
Privacy	0.975	0.975	± 0
Utility	0.850	0.834	+0.016
Latency (ms)	6713	7350	−637
Unnecessary upgrades	32.0% [20.8, 45.8]		

Table 14: Verification A/B on SynthPAI ($n = 50$, Wilson CIs). OFF reproduces the raw-confidence behaviour of Staab/Has/AgentStealth.

Risk-control context definitions

Table 22 lists the four action contexts used in §4.3. Each context is a tuple (recipient, purpose, action basis, policy transparency); these feed directly into the AMRSF multipliers M_{tr} , M_p , M_b , and M_T defined in Method §3.3. The likelihood L is held fixed at 0.7 across all contexts so that any variation in R is attributable to context, not to the LLM.

Per-context mean final level (transform-band)

Table 20 reports the mean final level after the verified-guesser adaptive loop, for samples in the transform region of each band context (Llama-3.2-3B, $n = 200$).

Risk-control qualitative cases (TAB, SynthPAI, PII-Masking)

Table 23 showcases 3 examples for anonymization in TAB, SynthPAI and PII-Masking-30k as well as risk score with decisions.

I LLM Evaluation Protocol

We pre-registered a three-rater by fifty-sample by eight-method evaluation study (1200 ratings) on SynthPAI subsamples, rating each (original, anonymized) pair on three 1–5 Likert dimensions: *Privacy* (inferability of personal attributes), *Utility* (meaning preservation), and *Fluency* (naturalness). Methods are shuffled and anonymized on the rater’s worksheet; raters work independently.

Method	TAB		SynthPAI	
	Priv	Util	Priv	Util
D. GuardianAgent	.941 _{.024}	.695 _{.012}	.962 _{.015}	.879 _{.034}
G. HaS	.925 _{.017}	.668 _{.020}	.797 _{.015}	.845 _{.009}
F. CONFAIDE	.899 _{.024}	.641 _{.009}	.843 _{.026}	.830 _{.020}
K. Pissarra	.881 _{.026}	.712 _{.019}	.790 _{.055}	.875 _{.023}
I. Staab’25	.667 _{.052}	.586 _{.048}	.986 _{.018} [†]	.272 _{.042}

Table 15: Four-seed variance ($n = 50$ per seed, mean with std as subscript, 5-level flagship). [†]Staab’s SynthPAI privacy still reflects 22% empty outputs. GuardianAgent is rank-1 privacy among methods with well-formed outputs at every seed.

Method	Llama-3.2-3B		Llama-3.1-8B	
	Priv	Util	Priv	Util
GuardianAgent (full)	0.975	0.855	1.000	0.863
Staab	0.975	0.225	0.950	0.700

Table 16: Backbone sensitivity (SynthPAI, $n = 50$). Privacy ranking preserved; Staab utility climbs at 8B but GuardianAgent already saturates at 3B.

We report inter-annotator agreement via Krippendorff’s α (target ≥ 0.68) and test pairwise significance with a paired bootstrap ($n=1000$ resamples, $p<0.05$). The eight configurations comprise three GuardianAgent variants and five baseline methods: Staab, Presidio, HaS, AgentStealth, and DP-Prompt.

LLM-rater simulation. As a proxy for the human study, we instantiate the same protocol with three Claude models acting as raters: Claude Opus 4.7, Claude Sonnet 4.6, and Claude Haiku 4.5. Each rater sees the (original, anonymized) pair and returns a strict-JSON Likert triple under a fixed system prompt. The privacy rubric explicitly penalizes destructive over-rewriting — a high Privacy score requires both that identifying details are removed *and* that the text remains usable; placeholder-only or content-free outputs are treated as destruction, not anonymization. Table 21 reports per-method consensus means (3-rater average) and paired-bootstrap p -values vs. GuardianAgent. To isolate rewriting style from the adaptive controller, GuardianAgent is evaluated under L1-L5 adaptive, L4 fixed and L5 fixed anonymization. Krippendorff’s α across the three Claude raters meets the pre-registered target of 0.68 on all three dimensions (privacy 0.68, utility 0.74, fluency 0.75), indicating substantial inter-rater agreement.

Class	SynthPAI	TAB
CLEAN_SUCCESS	65.5% [58.7, 71.7]	54.5% [47.6, 61.3]
CASCADE (L4 ×3)	16.5%	22.0%
PII_RETAINED	3.5%	0.5%
PARSE_FAIL	1.0%	0.0%

Table 17: Error analysis (Wilson 95% CIs on the main class). Dominant failure mode is hard PII-dense samples, not broken outputs.

J Computation resources

The primary LLM backbone is **Llama-3.2-3B-Instruct** (3.2B parameters), served via vLLM on a single NVIDIA H100 NVL (95 GB). Cross-model checks use **Llama-3.1-8B-Instruct** (8B) and **Qwen3-8B** (8B). Claude raters for the simulated human evaluation are Anthropic **Claude Opus 4.7**, **Sonnet 4.6**, and **Haiku 4.5**.

Compute budget (this paper): 2 NVIDIA H100 NVL + API

K Use of AI Assistants

We used a conversational AI assistant to (a) aid in writing and editing manuscript — specifically grammar polish and table layout; no scientific claim in this paper was generated by an AI assistant.

Method	TAB		SynthPAI	
	Priv	Util	Priv	Util
A. GuardianAgent	0.976	0.628	0.927	0.943
B. Staab'25	0.882	0.549	0.920	0.818
D. HaS'23	0.879	0.728	0.724	0.938
E. CONFAIDE'24	0.408	0.848	0.464	0.966
F. Pissarra'24	0.855	0.707	0.593	0.962

Table 18: Qwen3-8B backbone cross-check ($n! = 200$). GuardianAgent is rank-1 privacy on both corpora; CONFAIDE collapses. The five most private methods are shown.

Iter.	Events	S1 acc	Escalation
0	0	0.920	3.75%
1	500	0.970	2.00%
2	1000	0.993	0.25%
3	1500	0.998	0.25%
4	2000	1.000	0.25%

Table 19: Online bootstrapping of System 1 from System 2 teacher labels over a 2000-event OPP-115 stream; held-out 400-event test set. 24 teacher labels suffice to reduce escalation by 93% relative while pushing accuracy to perfect.

Context	ℓ_0	ℓ_{final}	$\Delta\ell$
T_{mild}	1.48	2.37	+0.89
T_{moderate}	2.52	2.93	+0.41
T_{strong}	3.33	3.64	+0.31
$T_{\text{near-deny}}$	3.98	4.42	+0.44

Table 20: Per-context mean final anonymization level after adaptive escalation. The verified guesser can escalate to L5 even from a low starting level when residual leakage remains, so $\ell_{\text{final}} \geq \ell_0(z)$. The largest $\Delta\ell$ occurs at T_{mild} , where the controller starts at L1 but escalation lifts the average toward L2–L3 on samples with residual surface-supported leakage.

Method	Priv $\uparrow$	Util $\uparrow$	Fluency $\uparrow$
D. Ours (L1–L5 adaptive)	3.95_{.78}	4.16_{.87}	4.20 _{.82}
D. Ours (L4 fixed)	3.74 _{.94}	3.91 _{.81}	4.22 _{.83}
D. Ours (L5 fixed)	4.12_{.73}	3.54 _{.88}	4.19 _{.72}
E. Presidio	3.17 _{.65}	3.64 _{.68}	3.38 _{.82}
G. HaS	2.93 _{.73}	2.99 _{1.00}	2.85 _{1.02}
H. DP-Prompt	3.19 _{.95}	3.90 _{.43}	4.27_{.67}
J. AgentStealth	3.71 _{.66}	3.36 _{.72}	3.68 _{.75}
I. Staab	2.08 _{1.40}	1.95 _{1.37}	1.97 _{1.45}
<i>Krippendorff's α (interval, 3 raters)</i>			
	0.68	0.74	0.75

Table 21: LLM evaluation on SynthPAI samples (1200 ratings). Three Claude models (Opus 4.7, Sonnet 4.6, Haiku 4.5) act as the three pre-registered raters. GuardianAgent is evaluated under L1-L5 adaptive, L4 fixed, and L5 fixed. † Paired-bootstrap two-sided $p < 0.05$ vs. Ours (L1–L5 adaptive) at $n=1000$ resamples.

Context	Recipient	Purpose	Action basis	Policy transparency
C_{low}	first_party	Functionality	user-initiated (paste)	full: categories, purposes, retention, legal basis, rights
C_{med}	service_provider	Analytics	default (passive)	partial: categories and purposes only
C_{high}	advertising	Profiling	background (script)	vague: short non-specific snippet
C_{ext}	data_broker	Advertising	background (script)	missing: no policy evidence

Table 22: Action contexts used in the risk-to-level study. Recipient feeds M_{tr} , purpose feeds M_p , basis feeds M_b , transparency feeds M_T . The categories themselves come from each sample’s annotated PII labels.

Context	R	Decision	ℓ_0	Final ℓ	Anonymized output (bold marks tokens introduced by the anonymizer)
<i>TAB</i> TAB_0025. Original: “The formal record of the judge’s order was not prepared by the parties until 26 October 1995. The applicants and Mrs Price applied to the Court of Appeal for leave to appeal on 1 November 1995. The appeal bundles were lodged on 4 October 1996. On 15 January 1997 the appeal was listed for hearing, and it was heard on 16 and 17 April 1997, on which last date the Court of Appeal gave judgment dismissing the appeal.”					
C_{low}	.108	allow	L1	L1	The formal record of the judge’s order was not prepared by the parties until 1995 . The applicants and a female applied to the Court for leave to appeal on 1995 . The appeal bundles were lodged on 1996 . On 1997 the appeal was listed for hearing, and it was heard on 1997 , on which last date the Court gave judgment dismissing the appeal.
C_{med}	.376	transform	L1	L1	Identical to C_{low} .
C_{high}	.910	deny	L4	L4	The formal record of the judge’s order was not prepared by the parties until a specific date . The applicants, including a female applicant , applied to a high court for leave to appeal on a specific date . The appeal bundles were lodged on a specific date . On a specific date , the appeal was listed for hearing, and it was heard on two specific dates , on which last date a high court gave judgment dismissing the appeal.
C_{ext}	.910	deny	L4	L4	Identical to C_{high} .
<i>SynthPAI</i> STAAB_0025. Original: “Ah reminiscing childhood, huh? Those were the days... I was completely enraptured by my Famicom. ... this game — Dragon Quest III. Eiichiro Nakatsu, the genius. ... That’s probably why I ended up a game developer now, haha. Nostalgic TIMESSSS!!”					
C_{low}	.108	allow	L1	L2	Ah reminiscing childhood, huh? Those were the days... I was completely enraptured by my Famicom. ... this game — Dragon Quest III. Eiichiro Nakatsu, the genius. ... That’s probably why I ended up a [PROFESSION] now, haha. Nostalgic TIMESSSS!!
C_{med}	.376	transform	L1	L2	Identical to C_{low} .
C_{high}	.910	deny	L4	L5	Ah reminiscing childhood, huh? Those were the days... I was completely enraptured by my Famicom. ... this game — Dragon Quest III. Eiichiro Nakatsu, the genius. ... That’s probably why I ended up in a creative field now, haha. Nostalgic TIMESSSS!!
C_{ext}	.910	deny	L4	L5	Identical to C_{high} .
<i>PII-Masking-300k</i> PII_41439A. Original: “Quality Assurance Feedback Thread. Subject: QA Review for Drug and Alcohol Education Course. Dear [Mayoress LORYA], We have completed a comprehensive Quality Assurance (QA) review of the Drug and Alcohol Education course delivered in our institution. We are pleased to provide you with detailed feedback on the performance of the course as well as individual feedback for each participant.”					
C_{low}	.108	allow	L1	L4	Quality Assurance Feedback Thread. Subject: QA Review for Drug and Alcohol Education Course. Dear [Institutional Leader] , We have completed a comprehensive Quality Assurance (QA) review of the Drug and Alcohol Education course delivered in our institution. We are pleased to provide you with detailed feedback on the course’s performance, as well as individual feedback for each participant.
C_{med}	.376	transform	L1	L4	Identical to C_{low} .
C_{high}	.910	deny	L4	L4	Identical to C_{low} (starting from L4 produces the same output without L1→L4 escalation rounds).
C_{ext}	.910	deny	L4	L4	Identical to C_{high} .

Table 23: Outputs for allow and deny rows are counterfactual anonymizer outputs; in deployment, allow passes the original text and deny blocks transmission. Qualitative risk-to-level examples on all three corpora. **Bold** tokens are spans introduced by the anonymizer, replacing original surface forms. **TAB:** the verified guesser is satisfied at L1 for low/medium contexts (final level matches ℓ_0); the L4 rewrite under high/extreme contexts replaces calendar dates and proper-noun courts with generic descriptions. **SynthPAI:** only the profession leaks at L1, so the guesser escalates one step (L1→L2) in low/medium contexts; under high/extreme contexts, the L5 semantic rewrite paraphrases the profession instead of inserting a tag. **PII-Masking:** the tagged recipient identifier is still recoverable at L1, so the verified guesser escalates all the way to L4 even under C_{low} ; high/extreme contexts start at L4 and converge to the same output. In every case, the *starting* level $\ell_0(R)$ shifts monotonically with action context, while the verified guesser additionally lifts the level whenever residual leakage is surface-supported.