OPERATIONALIZING DATA MINIMIZATION FOR PRIVACY-PRESERVING LLM PROMPTING

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid deployment of large language models (LLMs) in consumer applications has led to frequent exchanges of personal information. To obtain useful responses, users often share more than necessary, increasing privacy risks via memorization, context-based personalization, or security breaches. We present a framework to formally define and operationalize **data minimization**: for a given user prompt and response model, quantifying the least privacy-revealing disclosure that *maintains* utility, and propose a priority-queue tree search to locate this optimal point within a privacy-ordered transformation space. We evaluated the framework on four datasets spanning open-ended conversations (ShareGPT, WildChat) and knowledge-intensive tasks with single-ground-truth answers (CaseHold, MedQA), quantifying achievable data minimization with nine LLMs as the response model. Our results demonstrate that larger frontier LLMs can tolerate stronger data minimization while maintaining task quality than smaller open-source models (**85.7% redaction** for GPT-5 vs. **19.3%** for Qwen2.5-0.5B). By comparing with our search-derived benchmarks, we find that LLMs struggle to predict optimal data minimization directly, showing a bias toward abstraction that leads to oversharing. This suggests not just a privacy gap, but a capability gap: *models may lack awareness of what information they actually need to solve a task.*

1 INTRODUCTION

Users increasingly reveal sensitive personal information to large language model (LLM) applications (Mireshghallah et al., 2024a; Zhang et al., 2024), exposing themselves to privacy leaks via memorization, context-based personalization, or security breaches. Many share details believing it boosts task performance (Zhang et al., 2024), but this benefit is often illusory: people routinely overshare beyond what utility requires (Zhou et al., 2025). We ask a fundamental question: *What is the minimal information needed to maintain utility while preserving privacy?* This question is essential to quantify oversharing—that is, to compare actual disclosure against the true minimum.

Data minimization, defined as limiting the collection of personal information to what is necessary to accomplish a specified purpose, is a well-established privacy design pattern (Cavoukian et al., 2009) and is explicitly cited in numerous privacy regulations (e.g., GDPR (Parliament & Council, 2016)). Although considerable work has sought to mitigate the oversharing of sensitive information in LLM applications, few studies explicitly *formalize or quantify* this challenge from the perspective of data minimization. Existing approaches typically focus on detecting personal or sensitive disclosures and then apply redaction (e.g., "New York" → "[GEOLOCATION]") or abstraction (e.g., "New York" → "a city in the U.S.") (Dou et al., 2024; Zeng et al., 2025); related efforts develop heuristics to flag information types that are sensitive yet have low semantic relevance to the task (e.g., SSNs (Chowdhury et al., 2025)) or employ LLM-as-a-Judge to assess the relevance or importance of information to guide sanitization (Ma et al., 2025; Ngong et al., 2025). In this work, we introduce a framework that formally operationalizes data minimization for privacy-preserving LLM prompting, and present an algorithm that searches for the minimum privacy disclosure while preserving utility, thereby providing an oracle of data minimization for any prompt and target response generation model.

Figure 1 illustrates our framework with a running example. Our method can be viewed as a specialized tree search for data minimization. Starting from a root node that represents the most heavily sanitized prompt—capturing the globally most privacy-preserving formulation—we iteratively

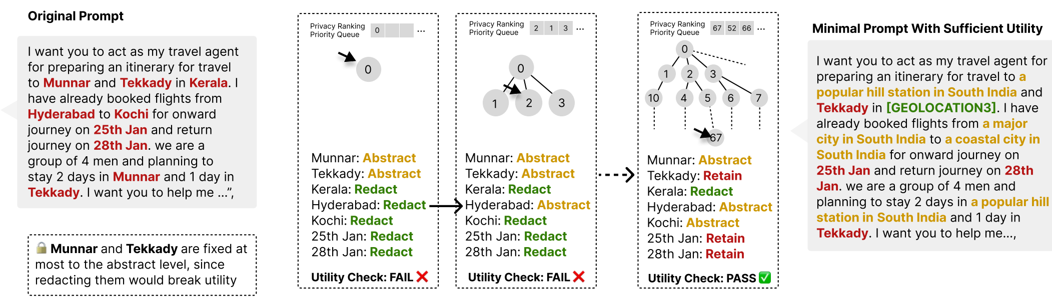

Figure 1: Framework Overview. We present a running example to demonstrate how we perform a tree search ranked by privacy variants, and a transformation that achieves data minimization.

expand the tree. Unlike classical depth-first or breadth-first search, we maintain a priority queue ordered by privacy sensitivity. At each step, we dequeue the least sensitive node, generate slightly more informative (and thus more privacy-revealing) variants as its children, and enqueue them. This process systematically explores the space of possible prompts in order of increasing privacy disclosure, enabling the identification of a minimally sufficient prompt that satisfies the target utility.

Our experimental results show that even under this utility-first constraint, there remains significant room for preserving privacy with data minimization—far exceeding the level of protection typically achieved in current practice. We observe that more powerful frontier models offer greater potential for data minimization than smaller, less capable ones. On open-ended real-world LLM prompts, gpt-5 shows the strongest removal with 85.7% REDACT and 8.6% ABSTRACT (only 5.7% RETAIN), while the smallest model (qwen2.5-0.5b) lags with 19.3% REDACT, 11.0% ABSTRACT, and 69.7% RETAIN.

By comparing with our oracles, we show that LLMs from small edge models to frontier reasoning models are poor predictors of data minimization, which bias towards ABSTRACT actions, leading to prevalent oversharing predictions. Together, these results demonstrate data minimization as a promising paradigm for addressing input privacy in LLM systems, while also revealing gaps in the popular LLM-as-a-Judge method for privacy-utility assessment tasks (Ma et al., 2025; Ngong et al., 2025). **This suggests not just a privacy gap, but a capability gap: models may lack awareness of what information they actually need to solve tasks**. We call for research to investigate the underlying causes of the varied levels of information "redundancy" across models, with the goal of developing robust prediction methods for effective on-device data minimization.

2 Background & Related Work

Theoretical & regulatory foundation. LLMs can expose memorized training data and personally identifiable information (PII) under adversarial prompting, motivating a shift toward minimizing user-side disclosure before inference rather than relying solely on post-hoc filtering. This imperative embodies the data minimization principle, a cornerstone of privacy laws and design guidelines. For example, data minimization is a pillar of the privacy by design framework (Cavoukian et al., 2009), a foundational and widely recognized regulatory framework central to modern data protection regimes such as GDPR Art. 5(1)(c), which limits processing to data necessary for a specified purpose (Parliament & Council, 2016).

User-led minimization for prompts. User-assisted tools help them manually sanitize inputs prior to submission (Zhou et al., 2025; Kan et al., 2023). However, these workflows hinge on subjective judgments of what "feels safe," offer *no guarantees of utility preservation*, and rarely include *attacker-based verification* of residual leakage. User studies on implicit inference further show people systematically *underestimate* what models can infer and often choose ineffective rewriting strategies (e.g., paraphrasing) (Wang et al., 2025). In contrast, we automate selection among {RETAIN, ABSTRACT, REDACT} in accordance with the data minimization principle by expanding a tree in increasing order of privacy disclosure, with a priority queue guiding the exploration based on this privacy order. We employed attacker LLMs tasked with type-wise and span-wise recovery of the redacted and abstracted information in the minimized prompts produced by our method, further verifying the limited recoverable signal and the efficacy of the minimization.

Utility-preserving minimization and prompt sanitizers. Prior input sanitization methods either do not consider utility (Dou et al., 2024), seek a balance between privacy and utility (Li et al., 2025b), or aim to maximize utility under a privacy constraint (e.g., a differential privacy budget Chowdhury et al. (2025)). Data minimization, representing a class of methods that optimize privacy under strict utility constraints, has received limited attention. A related line of work relies either on heuristics (e.g., detecting tokens whose format alone indicates sensitive content, such as SSNs Chowdhury et al. (2025)) or on LLM-as-a-Judge to assess how essential or relevant a piece of information is to the task, and then transforms the less essential and sensitive information to maintain utility (Ma et al., 2025; Ngong et al., 2025). However, we caution that it remains unclear to what extent LLM assessments align with the actual importance or necessity of the information, as this alignment depends not only on the semantic meaning of the information and the task but also on the capability of the target model. Our results further show that LLMs are poor predictors of data minimization, highlighting this gap.

Training-stage defenses (orthogonal). Differentially private (DP) training/fine-tuning (Abadi et al., 2016) and machine unlearning (Bărbulescu & Triantafillou, 2024) offer training-side protection against the downstream harms of oversharing caused by memorization during training. These approaches require access to model parameters and incur utility and compute costs, and they do not address other threat models to which oversharing is also vulnerable, including inference-stage leakage (Shao et al., 2025), data breaches (Theori Research, 2025; Meta Security Team, 2025; Gadget Review, 2025), or uninformed consents (Zhang et al., 2024; Fast Company, 2025). Our method is *black-box and pre-inference*: it operates solely on the *user input* and uses output-level utility checks, complementing these methods by remaining compatible with closed and rapidly evolving models, when fundamentally mitigating multiple threats through protection of the initial disclosure.

3 Data Minimization for Privacy-Preserving LLM Prompting

3.1 Problem Formulation

Let x be a user message and let $D = \{e_1, \ldots, e_n\}$ be a set of detected sensitive spans. Each span e_i can be transformed by an action a_i chosen from some finite action space A, forming an action vector $a = (a_1, \ldots, a_n)$. Applying a to x yields a transformed message $\tau(x; a)$. Given a target large language model $\mathcal{F}$, we seek a transformation that maximizes privacy while preserving downstream utility. Because placeholders or abstractions may later be replaced with their recovered context, the utility is evaluated *after* a context-recovery step $\mathcal{R}$ that reconstructs a usable output from $\mathcal{F}$:

$$\max_{a \in A^n} \ \mathrm{Priv}\big(\tau(x; a)\big) \quad \text{subject to} \quad \mathrm{Util}\big(\mathcal{R}(\mathcal{F}(\tau(x; a)))\big) \geq \gamma, \tag{1}$$

where

- Priv is any privacy metric (e.g., risk of sensitive-entity disclosure),
- Util is any downstream utility metric evaluated on the recovered output $\mathcal{R}(\mathcal{F}(\cdot))$,
- $\mathcal{R}$ is the context-recovery operator that replaces placeholders or abstractions with the appropriate recovered content, and
- γ is a minimum acceptable utility level.

This formulation is agnostic to the choice of action space, privacy/utility metrics, and search strategy.

3.2 Specific Instantiation

In this instantiation, we ground the generic formulation by defining a span-level action space $A = \{\text{RETAIN}, \text{ABSTRACT}, \text{REDACT}\}$, which we arrange as an ordinal hierarchy reflecting increasing privacy strength. Each detected sensitive span e_i is assigned one of these actions, inducing a space of possible variants guided by human preferences for privacy sensitivity. The algorithm searches this preference-ordered space to identify the most privacy-preserving variant while ensuring that the utility predicate yields an acceptable judgment. This construction provides the foundation for the formal definitions that follow.

Action Space. The action space is $A = \{\text{RETAIN}, \text{ ABSTRACT}, \text{ REDACT}\}$. These actions form an *ordinal lattice*, RETAIN $\prec$ ABSTRACT $\prec$ REDACT, encoding increasing privacy strength. The lattice is used to define one-step relaxations for the search procedure, and identify spans that cannot be modified without violating the utility constraint (Stage 1 of our search algorithm).

Utility Predicate. Let $y = \mathcal{F}(x)$ and $\tilde{y} = \mathcal{F}(\tau(x; \mathbf{a}))$. For open-ended tasks, placeholders/abstractions in $\tilde{y}$ are deterministically restored to $\tilde{y}^{\text{rb}}$ using the transformation map. A judge model then evaluates the pair $(y, \tilde{y}^{\text{rb}})$ under a fixed rubric to verify that the transformation does not degrade task performance, returning `pass` or `fail`. For tasks with fixed ground truths, utility is `pass` iff $\mathcal{F}(\tau(x; \mathbf{a}))$ is correct under the task's scoring rule (e.g., exact match or multiple-choice accuracy). The only criterion for accepting a candidate is the utility predicate UTIL returns `pass`.

We examined how sensitive the utility predicate is to small relaxations of the threshold γ. To test whether users can perceive such utility reductions, we conducted a user study (see Appendix F) comparing outputs produced under different γ settings. The results show that even minor relaxations of γ lead to noticeable quality degradation from a user's perspective. This supports our choice of a strict pass-fail utility predicate that requires preserving the original utility without degradation.

Privacy Comparator. To define a structured search space over privacy transformations, we introduce a pairwise privacy comparator $\mathcal{C} : (x, \tau_A, \tau_B) \longmapsto \{\tau_A, \tau_B, \text{ SAME}\}$. Given two variants of the *same* source message, it returns which is more privacy-preserving (or SAME).

Unlike a partial order, this relation is not assumed to be transitive or total, reflecting the empirical reality that human privacy preferences may exhibit intransitivities or context-dependent judgments. Our algorithm leverages this relation as an ordering signal, treating it as an oracle for guided search without requiring formal lattice properties.

4 Algorithm and Implementation

This section presents both the algorithmic procedure and the practical implementation of our framework. The algorithm specifies a two-stage search over the privacy-ordered action space, and the implementation focuses on instantiating the privacy comparator to align with human preferences. Together, they define the end-to-end system used in our experiments.

4.1 Algorithm: Freeze-Then-Search

Stage 1: Freeze Inflexible Entities. For each $e \in D$, probe REDACT(e) and ABSTRACT(e) in isolation while keeping all other entities RETAIN. If both probes cause utility to fail, mark e as *frozen* (forced RETAIN thereafter). Let $D' \subseteq D$ be the non-frozen entities with $n' = |D'|$; only D' participates in Stage 2. This step both preserves utility invariants and reduces the branching factor.

Stage 2: Privacy-Comparator Priority-Queue Tree Search. The tree search begins at a root node obtained by applying to each $e \in D'$ the most privacy-preserving transformation allowed by Stage 1. Each node encodes a transformation action vector a and its corresponding transformed message $\tau(x; a)$. For any notes, child nodes are generated by relaxing exactly one action (e.g., REDACT $\rightarrow$ ABSTRACT; ABSTRACT $\rightarrow$ RETAIN). The tree is traversed in order of decreasing privacy, guided by a priority queue that uses $\mathcal{C}$ as the comparator. Ties (SAME) are broken by stable insertion order. The complete search procedure is given in Algorithm 1.

The procedure returns the *first* action profile $\mathbf{a}$ that satisfies the utility predicate. We record (i) the transformed input $\tau(x; \mathbf{a})$; (ii) the Stage 1 freeze set D' (entities forced to RETAIN); (iii) the per-entity action map. If no candidate passes, we return $\text{RETAIN}^{|D|}$.

Complexity. Stage 2 explores at most $|\mathcal{M}| = 3^{n'}$ action profiles on the non-frozen coordinates. If T candidates are expanded, a binary-heap implementation requires at most $C \le c\,T \log T$ pairwise comparisons (many avoided by caching). With average per-call latencies $t_{\mathcal{C}}$ and t_{UTIL} for comparator and utility respectively, $\text{Time} \lesssim c\,T \log T \cdot t_{\mathcal{C}} + T \cdot t_{\text{util}}$.

Algorithm 1: Privacy-Comparator Priority Queue Tree Search (Stage 2)

Input: message x; non-frozen entities D'; utility predicate U; comparator $\mathcal{C}$
Output: first passing action profile $\mathbf{a}$
Initialize $\mathbf{a}_0$: for $e \in D'$, set REDACT unless it failed in Stage 1 (then ABSTRACT); for $e \notin D'$, set RETAIN;
$Q \leftarrow$ comparator-based priority queue seeded with $\mathbf{a}_0$ (ties: stable order);
$V \leftarrow \emptyset$; `// visited`
while Q *not empty* **do**
  $\mathbf{a} \leftarrow Q.\text{pop}()$; **if** $\mathbf{a} \in V$ **then**
    continue
  $V \leftarrow V \cup \{\mathbf{a}\}$;
  if $U\big(\mathcal{F}(x),\, \mathcal{F}(\tau(x; \mathbf{a}))\big) = \mathit{pass}$ **then**
    return a
  foreach $e \in D'$ *with* $a_e \in \{\text{REDACT}, \text{ABSTRACT}\}$ **do**
    $\mathbf{a}' \leftarrow$ degrade a_e by one step (REDACT$\rightarrow$ABSTRACT or ABSTRACT$\rightarrow$RETAIN);
    if $\mathbf{a}' \notin V$ **then**
      push $\mathbf{a}'$ into Q
return RETAIN$^{|D|}$ `// fallback`

4.2 Implementation

Privacy Transformations and Utility Check. For each prompt we fix detected PII spans D and a per-entity variants map (e.g., New York City and NYC) detected and clustered by GPT-4o; identical REDACT/ABSTRACT mappings and GPT-4o-generated abstractions are used across all models (App. D). We implement the span-level privacy transformation actions with a deterministic rewriter that (i) applies per-entity actions $a_i \in \{\text{RETAIN}, \text{ABSTRACT}, \text{REDACT}\}$ to produce $\tau(x; \mathbf{a})$ and a replacement map, and (ii) performs strict replace-back on model outputs for evaluation (Sec. 3.2). For utility, GPT-4o acts as judge (App. E): fixed-ground-truth tasks use the task's official scorer on $\mathcal{F}(\tau(x; \mathbf{a}))$; open-ended tasks are judged once on $(y, \text{restore}(\tilde{y}))$; single-answer QA runs k=5 independent decodes with early stop at the first mismatch, passing only if all k are correct.

Privacy Comparator. We collect human ground-truth labels on 150 A/B pairs sampled from a PII-rich subset of the ShareGPT dataset (RyokoAI, 2023), with each pair annotated by at least five annotators. Independently, we create 4,840 additional pairs and obtain teacher labels from a strong zero-shot judge (OpenAI O3) for supervised LoRA finetuning (Hu et al., 2022), resulting in a latency-optimized comparator (finetuned Qwen2.5-7B-Instruct; hyperparameters in App. B) Compared with the human labels, the distilled comparator achieves **71%** overall and **89%** in high-human-consensus items (≥ 0.8) at **0.31**s/decision—yielding a $> 20\times$ speedup vs. the zero-shot judges with comparable high-consensus accuracy (Table 1). This choice materially reduces the $c\,T \log T \cdot t_{\mathcal{C}}$ term in §14 and enables practical Stage 2 search. Consensus among the 150 human-labeled pairs varies substantially: 73 items reach consensus ≥ 0.8 and 121 reach ≥ 0.6, with only a small subset achieving full agreement. Comparator accuracy improves with higher consensus, rising from 71% overall to 77% at ≥ 0.6 and 89% at ≥ 0.8.

Comparator	Accuracy (All)	Acc. @ consensus ≥ 0.8	Latency (s)
o1 (zero-shot)	70%	90%	8.05
o3 (zero-shot)	70%	89%	6.37
o3-mini (zero-shot)	69%	88%	4.32
`Qwen2.5-7B-Instruct (finetuned)`	71%	89%	0.31

Table 1: Privacy comparator alignment with human judgments and per-decision latency.

Utility Evaluator. For open-ended tasks, where utility cannot be measured deterministically, we validate GPT-4o as the utility judge using samples drawn from the oracle's search trace. We constructed 150 evaluation pairs, balanced between GPT-4o PASS and FAIL decisions, and collected judgments from 75 human annotators (five per item) on whether utility was preserved (ACCEPT) or degraded (REJECT). Agreement between GPT-4o and humans increases with consensus strength, rising from approximately 0.69 at a 0.6 consensus threshold to approximately 0.94 under full agreement. This pattern parallels that of the privacy comparator and supports GPT-4o's reliability in the high-consensus regime where the utility predicate is most informative.

5 Experimental Details

5.1 Datasets and Prefilter

We sample test prompts from four datasets spanning open-ended and closed-ended tasks: ShareGPT (RyokoAI, 2023) (open-ended; 176 messages), WildChat (Zhao et al., 2024) (open-ended; 139), MedQA (Jin et al., 2020) (medical MCQ; 108), and CaseHOLD (Zheng et al., 2021) (legal MCQ; 110). All prompts contain PIIs (open-ended: ≥ 3; close-ended: ≥ 1).

For closed-ended datasets, we ensure that all test models can correctly answer the selected questions five times, so that any further accuracy drop can be attributed to reduced disclosure rather than intrinsic task difficulty. Open-ended datasets are prefiltered to only include PII-rich English text with a clear task. Detailed curation criteria are given in App. C.

5.2 Model Selection

We evaluate *nine* target models: *gpt-4.1-nano*, *gpt-4.1*, *gpt-5*, *claude-3-7-sonnet-20250219* (extended thinking disabled), *claude-sonnet-4-20250514* (extended thinking disabled), *lgai/exaone-deep-32b*, *mistralai/mistral-small-3.1-24b-instruct*, *qwen/qwen2.5-7b-instruct*, and *qwen/qwen2.5-0.5b-instruct*. This set covers a wide range of capacity model families, from frontier closed-weight models to small, open models suitable for on-device deployment. Two targets expose *reasoning modes* and are run with their default settings: *gpt-5* (default reasoning profile; `reasoning_effort=medium`) and *lgai_exaone-deep-32b* (provider default reasoning mode). All other targets are instruction-tuned chat models.

5.3 Experiment I: Establishing Data Minimization Oracles

We applied our framework to search data minimization, using the nine target models as the response-generation model $\mathcal{F}$ on prompts sampled from the four datasets. We report data minimization results as the optimal percentage of REDACT/ABSTRACT/RETAIN actions under the utility constraint.

To verify that minimization *robustly* reduces recoverability of masked information (redacted or abstracted) from the message itself, we run two black-box adversarial audits that attempt to simulate *on-text* inference by an adversary (Staab et al., 2024). **Type-wise recovery:** Given the text and the *set of types that were marked during minimization*, the attacker must output up to three *verbatim* candidates per requested type with confidences, relying only on the given text. We evaluate the same attacker on both the original input x and the minimized input $\tilde{x}$ with an identical type set. For each type, we compute Hit@1/Hit@3 against the corresponding gold strings. **Span-wise recovery:** Given the minimized text $\tilde{x}$ and the list of replacement strings actually inserted by our pipeline (e.g., `[NAME1]` or abstraction phrases), the attacker must, for *each* span, return a single guess of its original string or ``Unknown'' with confidence 0 if it cannot be recovered from this message alone. We use two LLMs different from the nine target test models as attackers: one open-weight model (`meta-llama/llama-3.1-70b-instruct`) and one closed-weight model (`google/gemini-flash-1.5`)

5.4 Experiment II: Benchmarking Zero-Shot LLM Data Minimization Predictors

With the oracles in place, we evaluate the selected models in the *prediction* setting: given an input, the model must directly choose an action from {RETAIN, ABSTRACT, REDACT} for each detected span to produce the most privacy-preserving variant while preserving utility, **without comparator guidance, search, or any in-loop utility judge**.

The prompt provides the message, span types, span variants, and the replacement strings that would be applied if chosen. We parse the model output into an action map; invalid actions are repaired with a schema-only prompt, and undecided spans are marked and excluded from conditioned ratios.

For each item i and predictor model m, we pair the oracle minimized prompt $\tilde{x}_i^\star$ with the predicted one $\tilde{x}_i^{(m)}$ to evalute with the same **pairwise sensitivity comparator** and **utility predicate** as in the search process. We classify (item, m) into four disjoint categories: *Overshare* if prediction leaks more privacy than oracle), *Undershare+Fail* if prediction is more protective but fails utility, *Undershare+Pass* if prediction is more protective and passes utility. *Fit* if prediction ties the oracle on privacy and passes utility. The first two categories are considered unsuccessful minimization, whereas the latter two represent successful minimization.

6 RESULTS

6.1 DATA MINIMIZATION ORACLE

Our minimization oracles show frontier models achieve the most privacy protection without violating the utility constraint (Table 2). On open-ended task prompts, *gpt-5* achieves the most aggressive removal—**85.7%** REDACT and **8.6%** ABSTRACT (only 5.7% RETAIN)—while the smallest model (*qwen2.5-0.5b*) sits at the bottom with **19.3%** REDACT and **11.0%** ABSTRACT (69.7% RETAIN). Closed-ended tasks admit even more minimization: *gpt-4.1* tops the board at **98.0%** REDACT and **1.0%** ABSTRACT (1.0% RETAIN), whereas *qwen2.5-0.5b* again trails with **32.1%** REDACT and **11.7%** ABSTRACT. The scatterplot in Fig. 2 shows frontier models clustered near the $x+y=1$ band, confirming that very little PII must be retained to preserve utility.

Overall, minimization is *redaction-heavy*: abstraction stays small (typically 1–12%), indicating that simply deleting sensitive spans is usually sufficient for the utility constraint. Smaller models accept far less minimization in both settings, which is acceptable in practice because they are more feasible to be deployed on-device, posing lower leakage risks. A cross-model Jaccard analysis (App. I) further shows that, despite differences in the exact minimized prompts, redaction decisions are highly consistent across model families. The majority of cross-model variation arises instead from the much smaller abstraction set, which both explains the larger fluctuations observed in abstraction and suggests that the core redactions transfer well across models.

Response Generation Model	**Open-ended**			**Closed-ended**		
	Redact ↑	**Abstract ↑**	**Retain ↓**	**Redact ↑**	**Abstract ↑**	**Retain ↓**
gpt-5	**85.7%**	**8.6%**	**5.7%**	97.1%	1.8%	1.1%
gpt-4.1	82.6%	9.9%	7.6%	**98.0%**	**1.0%**	**1.0%**
gpt-4.1-nano	79.6%	10.0%	10.5%	91.3%	2.0%	6.7%
claude-sonnet-4-20250514†	74.8%	11.2%	14.0%	97.2%	1.9%	0.9%
claude-3-7-sonnet-20250219†	77.5%	10.6%	11.9%	79.5%	10.1%	10.4%
lgai_exaone-deep-32b	60.4%	17.4%	22.2%	75.0%	10.2%	14.7%
mistral-small-3.1-24b-instruct	75.3%	12.5%	12.2%	96.4%	1.7%	1.9%
qwen2.5-7b-instruct	69.9%	12.0%	18.1%	91.7%	4.6%	3.7%
qwen2.5-0.5b-instruct	19.3%	11.0%	69.7%	32.1%	11.7%	56.2%

Table 2: Optimal percentage of REDACT, ABSTRACT, and RETAIN actions for open-ended (ShareGPT, WildChat) and closed-ended (MedQA, CaseHold) task prompts across nine models. ↑ indicates that higher is better, and ↓ indicates that lower is better. † Extended thinking disabled.

Span-wise Recovery. Pooling across target models and grouping spans by action (Table 3), **abstraction** consistently yields higher overall recovery than **redaction** on every dataset: the correct-recovery rate p_{corr} ranges **5.6–14.9%** for ABSTRACT versus only **2.7–7.7%** for REDACT. Importantly, the *absolute* rates are low across the board (all $p_{\text{corr}} < 0.15$, with REDACT ≤ 0.077), indicating that on-text inference

Table 3: Span-wise recovery pooled across target models: p_{corr} by action across (rows) datasets (columns).

Action	CaseHOLD	MedQA	ShareGPT	WildChat
abstract	0.092	0.056	0.149	0.119
redact	0.050	0.027	0.051	0.077

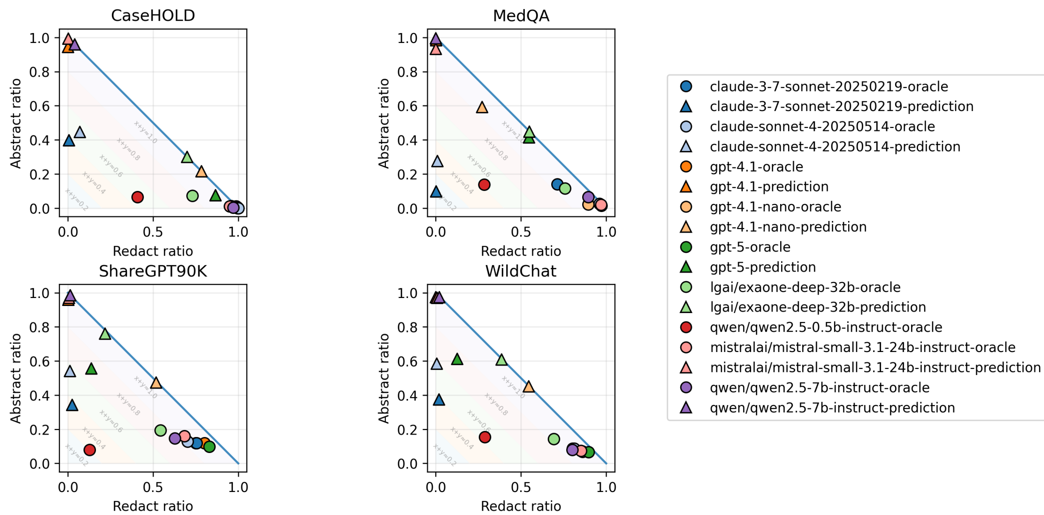

Figure 2: Oracle vs. Prediction REDACT and ABSTRACT Ratio.

is generally difficult under our setup. The separation is larger on open-ended data than on closed-ended data, suggesting that open-domain context leaves more clues. Redaction is more robust to on-text inference than abstraction—attackers both *attempt less* and *succeed less* after REDACT—and overall recovery remains low, reinforcing a *redact-first* policy when minimizing leakage, especially for open-ended inputs. A parallel span-level evaluation with GPT-5 as the attacker on its own leads to the same conclusion. Across datasets, correct-recovery rates remain low for both abstraction and redaction spans, and masked spans are overwhelmingly labeled as UNKNOWN. Full results are reported in Appendix J.2. Together, these findings show that GPT-5 is unable to reconstruct the removed private information even when attacking its own oracle-minimized prompts.

Type-wise Recovery (original vs. masked). Aggregating by entity type, masking causes a sharp drop in recoverability relative to the original text. For example on WILDCHAT (Hit@1, %), NAME falls from 90.3 to 0.0, GEOLOCATION from 89.8 to 2.2, OCCUPATION from 85.4 to 8.0, and AFFILIATION from 83.0 to 1.9; other datasets show the same pattern (Appendix J.1). Hit@3 mirrors these trends across types. In short, masking severely limits type-wise recovery. Consistent with the span-level results, a parallel test using GPT-5 as the attacker on its own minimized prompts shows the same pattern: masked Hit@1 for every type stays in the low single digits while the corresponding original values are often near the top of the scale, indicating that GPT-5 does not infer the removed PII.

Taken together, the span-wise and type-wise recovery checks confirm that our search-based data minimization method effectively strips sensitive information from prompts and prevents that information from being inferred indirectly from the remaining context.

6.2 PREDICTION VS. ORACLE

As shown in Fig. 3, single-pass predictions are generally less privacy-preserving than the `gpt-5` oracle—*Overshare* dominates across tasks—indicating that these direct predictions without comparator-guided search tend to under-protect privacy with frontier models which are most widely used and vulnerable to more privacy risks. Items counted as *Undershare+FAIL* reflect attempts to push masking beyond the oracle that break task utility. A meaningful slice—especially on open-ended datasets—falls into *Undershare+PASS*, signaling headroom to further tighten the oracle's comparator priorities or stop rule. The *Fit* mass (privacy tie + utility pass) is small, suggesting the prediction rarely sits close to a task-wise privacy/utility frontier. Oracles are harder to surpass in the close-ended, answer-verifiable tasks (MedQA is near-all Overshare, while CaseHOLD still shows non-trivial Undershare+PASS and Fit). Minor stochasticity in `gpt-5` decoding is mitigated via replace-back, and k=5 repetition on verifiable tasks.

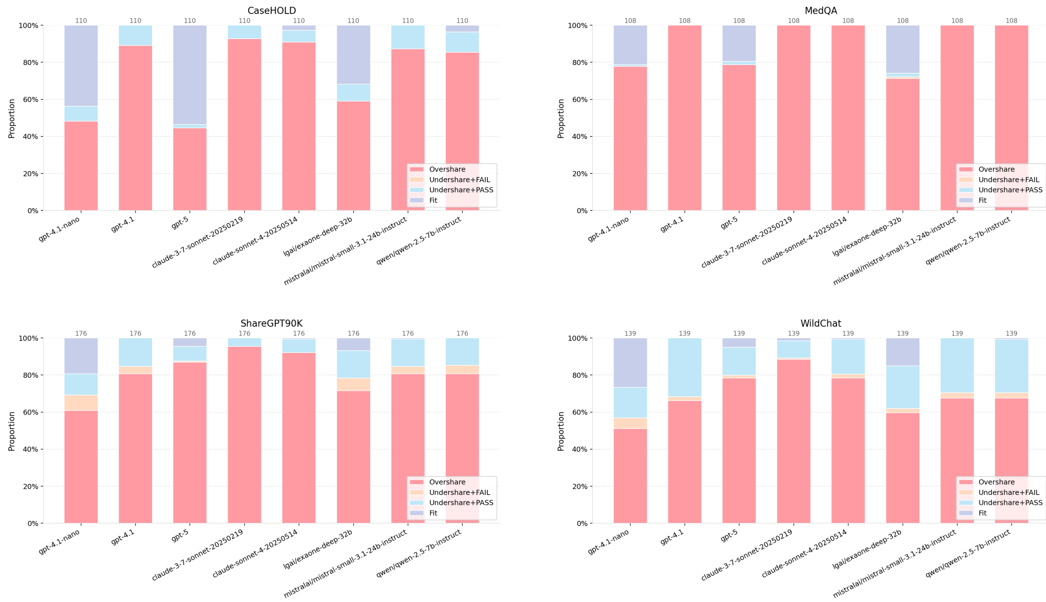

Figure 3: **Prediction vs. oracle minimization across datasets.** Each panel shows per-model stacked proportions that sum to 1. Outcomes are interpreted *relative to the* `gpt-5` *oracle* using our privacy comparator (Sec. 3.2) and utility predicate (Sec. 5): *Overshare*—the prediction disclosure is *less privacy-preserving* than the oracle; *Undershare+FAIL*—the prediction hides more but fails the utility check; *Undershare+PASS*—the prediction hides more and passes utility; and *Fit*—the prediction ties the oracle on privacy and passes utility.

Prediction bias toward ABSTRACT. In single-pass predictions, models consistently favor ABSTRACT over REDACT, showing an *abstraction-first* default on everyday user prompts (e.g., trip planning). Because ABSTRACT is less privacy-preserving in our setup, choosing it when REDACT would still retain utility implies unnecessary disclosure. This tendency persists even when instructions explicitly indicate that the protection strengths of REDACT is higher than ABSTRACT; and it contrasts with the oracles that cluster in the high-REDACT/low-ABSTRACT regime (cf. Fig. 2). We also tested whether the abstraction preference arises from our prompt design by ablating the minimization-order instruction. As detailed in App. G.4, removing either the "prefer stronger" clause or the entire minimization order line leaves model behavior nearly unchanged: all the selected models for this further test (GPT-5, Mistral-24B, and Qwen2.5-7B) still strongly prefer ABSTRACT, indicating that the bias is model-internal rather than prompt-induced.

Ablation by model family. Results show stable, high-level biases as illustrated by the prediction-side clusters in Fig. 2. **Mistral/Qwen/GPT-4.1** default to an *abstract-first* policy across datasets—even for *structured* identifiers—e.g., on ShareGPT and WildChat they abstract nearly all `URL`/`EMAIL`/`ID_NUMBER` spans with $\leq$1–2% redact and non-trivial retain on soft context like `GEOLOCATION`/`TIME`. **Claude** adds a pronounced RETAIN tail on open-ended prompts (large fractions of `GEOLOCATION`, `TIME`, `AFFILIATION` kept), with little redaction. By contrast, the two reasoning models **GPT-5** and **Exaone** are the only ones that *consistently redact* high-precision types: on closed-ended CaseHOLD/MedQA they heavily redact `NAME`/`TIME`/`GEOLOCATION`, and on open-ended chats they are far more willing than other families to redact `URL`/`EMAIL`/`PHONE_NUMBER`.

For completeness, we also observe that fully masking all detected PIIs, as would occur in a simple NER-based redaction, often breaks utility. Together with the oversharing behavior of single-pass predictions, this suggests that neither extreme is adequate, and an oracle is needed to determine how much masking each model can tolerate.

7 Conclusion and Discussion

We present a framework that formally defines and operationalizes data minimization in LLM prompting: for a given user prompt and response model, it quantifies the minimal privacy-revealing disclosure required to maintain utility. Our results show that data minimization offers a significant optimization space for reducing privacy exposure without compromising task performance, particularly for larger and more capable language models. However, we find that directly predicting this minimal disclosure is challenging, even for frontier models. This work lays the groundwork for research on quantifying data minimization and robust prediction methods, fostering both fundamental machine learning advances and interdisciplinary research in human-AI interaction.

Novel Paradigm of Privacy-Preserving LLM Interactions. We show that the more capable the model is, the more feasible data minimization becomes. This result shows that data minimization is a promising approach to addressing excessive disclosure problems in user interactions with LLM systems, as users tend to trade privacy for utility and therefore often choose frontier models hosted on the cloud for sensitive tasks despite privacy concerns (Zhang et al., 2024). The variances of data minimization across datasets and models suggest that model-specific predictors are needed, and we advocate that LLM providers include these as part of the released model package. Such predictors naturally align with an emerging line of work that explores a dual-model management approach: using small edge models for data-minimization-guided local sanitization before sharing data with the remote model (Li et al., 2025b; Zhou et al., 2025; Zhang et al., 2025; Chowdhury et al., 2025). Beyond these observations, our results also clarify the technical role of the oracle within this workflow. The oracle procedure identifies the upper bound of data minimization a target model can tolerate while preserving utility, providing high quality supervision for learning practical sanitization policies. This supervision can train or distill a small predictor that performs single pass span level decisions locally, complementing the dual model management approach described above. This establishes a natural path toward future on-device predictors that give users full control over the flow of private data before any interaction with a remote model.

LLM Capabilities for Privacy Tasks. We evaluate LLM capabilities on two novel privacy tasks: data minimization prediction and privacy sensitivity ranking (by the privacy comparator), extending prior work on using LLMs for PII detection and context-aware privacy judgments (Mireshghallah et al., 2024b; Shao et al., 2025; Li et al., 2025a). We find that data minimization prediction remains challenging for current state-of-the-art models. For the privacy sensitivity ranking task, we found that off-the-shelf reasoning models (e.g., GPT-o1, o3, and o3-mini) perform better than non-reasoning models (e.g., GPT-4o). Future research should further account for individual preference differences, as our results show that in over half of cases the five human raters reached a consensus score below 0.8. A failure case analysis of the best-performing models reveals where misalignment still occurs. In these cases, humans often choose “SAME,” while models prefer “A” or “B,” reflecting different thresholds for saliency: models overemphasize subtle distinctions that seem significant to them but are imperceptible or irrelevant to humans. Moreover, models tend to overvalue specificity and do not align with humans on how the specificity of certain data types corresponds to sensitivity (e.g., assigning more weight to time or date information than to names).

Interpretation Methods for “What is Necessary.” Foundational understanding of what information or tokens are necessary is still required to explain the variance observed in data minimization oracles across models and datasets. Current methods can reveal what information is used at inference (Vig et al., 2020), but determining what is truly necessary remains an open research frontier. In addition, the potential impact of test set contamination (Oren et al., 2024) should be carefully taken into consideration in future investigations.

Ethics Statement

All datasets are publicly available under their respective terms; we do not crawl private sources. All human-subjects studies have been approved by our institution’s IRB. Our human evaluation collects no PII of the human raters. Annotators only state preferences over sanitized replacements; no demographics are recorded and no unanonymized content is shown. Residual re-identification and misuse as obfuscation are potential risks; we mitigate them by favoring REDACT when utility allows, auditing with recovery attacks, and releasing only sanitized data and evaluation scripts.

REPRODUCIBILITY STATEMENT

We release an anonymous repository at anonymous GitHub repository. The pipeline code used in the oracle experiment (Section 5.3) is in `run_pipeline.py`; however, due to anonymization, our trained privacy comparator is hosted on a private cloud and its model ID cannot be disclosed, so an end-to-end run requires plugging in an alternative comparator. The folder `Prefiltered datasets` corresponds to the four test datasets described in Section 5.1. The file `human_annotation_vs_o3mini.jsonl` contains human annotator tallies (`tally`) and `o3mini` judgments used to (i) select the best teacher model, (ii) use the teacher to generate large input sets, and (iii) train the privacy comparator (Section 4.2); we also use the same human annotations to evaluate the comparator's accuracy. Setup notes, and example commands are provided in the repository README.

REFERENCES

Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016. doi: 10.1145/2976749.2978318. URL https://dl.acm.org/doi/10.1145/2976749.2978318.

George-Octavian Bărbulescu and Peter Triantafillou. To each (Textual sequence) its own: Improving memorized-data unlearning in large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3003–3023. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/barbulescu24a.html.

Ann Cavoukian et al. Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada*, 5(2009):12, 2009.

Amrita Roy Chowdhury, David Glukhov, Divyam Anshumaan, Prasad Chalasani, Nicolas Papernot, Somesh Jha, and Mihir Bellare. Pr$\epsilon\epsilon$mpt: Sanitizing sensitive prompts for llms. *arXiv preprint arXiv:2504.05147*, 2025. URL https://doi.org/10.48550/arXiv.2504.05147.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. Reducing privacy risks in online self-disclosures with language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13732–13754, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.741. URL https://aclanthology.org/2024.acl-long.741/.

Fast Company. Google is indexing conversations with chatgpt. https://www.fastcompany.com/91376687/google-indexing-chatgpt-conversations, 2025.

Gadget Review. Grok's privacy disaster: 370,000 ai conversations exposed on google. Technical report, 2025.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020. URL https://arxiv.org/abs/2009.13081.

Zhigang Kan, Linbo Qiao, Hao Yu, Liwen Peng, Yifu Gao, and Dongsheng Li. Protecting user privacy in remote conversational systems: A privacy-preserving framework based on text sanitization. *arXiv preprint arXiv:2306.08223*, 2023. URL https://arxiv.org/abs/2306.08223.

Haoran Li, Wenbin Hu, Huihao Jing, Yulin Chen, Qi Hu, Sirui Han, Tianshu Chu, Peizhao Hu, and Yangqiu Song. PrivaCI-bench: Evaluating privacy with contextual integrity and legal compliance. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10544–10559, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.518. URL https://aclanthology.org/2025.acl-long.518/.

Siyan Li, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. PAPILLON: Privacy preservation from Internet-based and local language model ensembles. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3371–3390, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.173. URL https://aclanthology.org/2025.naacl-long.173/.

Hongru Ma, Wenpeng Lu, Yanjie Liang, Tianyi Wang, Qi Zhang, Yingjie Zhu, and Jiasheng Si. Alsa: Context-sensitive prompt privacy preservation in large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 2042–2053, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736840. URL https://doi.org/10.1145/3711896.3736840.

Meta Security Team. Meta ai chatbot security flaw exposes user conversations, July 2025. Security Advisory.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. Trust no bot: Discovering personal disclosures in human-LLM conversations in the wild. In *First Conference on Language Modeling*, 2024a. URL https://openreview.net/forum?id=tIpWtMYkzU.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. In *International Conference on Learning Representations (ICLR)*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/08305d8b2ddab98932c163ea73df065f-Abstract-Conference.html. See also arXiv:2310.17884.

Ivoline C. Ngong, Swanand Ravindra Kadhe, Hao Wang, Keerthiram Murugesan, Justin D. Weisz, Amit Dhurandhar, and Karthikeyan Natesan Ramamurthy. Protecting users from themselves: Safeguarding contextual privacy in interactions with conversational agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26196–26220, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1343. URL https://aclanthology.org/2025.findings-acl.1343/.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KS8mIvetg2.

European Parliament and Council. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016.

RyokoAI. Sharegpt 90k conversations. https://huggingface.co/datasets/RyokoAI/ShareGPT52K, 2023. Accessed: 2025-09-24.

Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. Privacylens: evaluating privacy norm awareness of language models in action. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NeurIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=kmn0BhQk7p. Spotlight.

Theori Research. Deepseek security, privacy, and governance: Hidden risks in open-source ai. Technical report, Theori Inc., January 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Synthia Wang, Sai Teja Peddinti, Nina Taft, and Nick Feamster. Beyond PII: How users attempt to estimate and mitigate implicit LLM inference. *arXiv preprint arXiv:2509.12152*, 2025. URL https://arxiv.org/abs/2509.12152.

Hang Zeng, Xiangyu Liu, Yong Hu, Chaoyue Niu, Fan Wu, Shaojie Tang, and Guihai Chen. Automated privacy information annotation in large language model interactions. *arXiv preprint arXiv:2505.20910*, 2025. URL https://doi.org/10.48550/arXiv.2505.20910.

Juhua Zhang, Zhiliang Tian, Minghang Zhu, Yiping Song, Taishu Sheng, Siyi Yang, Qiunan Du, Xinwang Liu, Minlie Huang, and Dongsheng Li. DYNTEXT: Semantic-aware dynamic text sanitization for privacy-preserving LLM inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20243–20255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1038. URL https://aclanthology.org/2025.findings-acl.1038/.

Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. "it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642385. URL https://doi.org/10.1145/3613904.3642385.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024. URL https://arxiv.org/abs/2405.01470.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. *arXiv preprint arXiv:2104.08671*, 2021. URL https://arxiv.org/abs/2104.08671.

Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 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 (CHI '25)*. ACM, 2025. doi: 10.1145/3706598.3713701. URL https://arxiv.org/abs/2410.11876. Also available as arXiv:2410.11876 (v3, Feb 2025).

A LLM Usage

Outside of minor manuscript assistance, all LLM usage in this work was solely for running the experiments reported in the main text (PII detection/adjudication, abstraction-term synthesis, prediction runs, and adversarial audits), with model choices and procedures fully specified in Sec. 5

Manuscript assistance. We used LLMs only for light language polishing and LaTeX table formatting (e.g., caption phrasing, column alignment). No technical content or analysis was delegated, and all edits were author-verified.

B PRIVACY COMPARATOR TRAINING HYPERPARAMETERS

Qwen2.5-7B-Instruct (distilled final). learning_rate $= 1\text{e}{-4}$; epochs $= 2$; LoRA rank $= 8$; context length $= 2048$; global batch size $= 2048$ (Fireworks AI). On human-annotated evals (consensus ≥ 0.8), accuracy $= 89.04\%$.

C DATASET CURATION

Close-ended Dataset (MedQA and CaseHold) Each item is answered by all nine target models with five runs at temperature 0 on the original (unmasked) input; we keep an item only if *every* model is correct in *all five* runs. This ensures that any accuracy drop under minimization is attributable to masking rather than intrinsic task difficulty.

Open-ended Dataset (ShareGPT and WildChat) We curate a high-quality subset to study how far inputs can be minimized before utility degrades. Specifically, we (i) select PII-rich messages using Microsoft Presidio, requiring at least three *distinct* PII categories; (ii) retain primarily English content via fastText; (iii) remove near-duplicates using MiniLM embeddings; (iv) exclude code-heavy or prefixed web dumps and templated/meta prompts lacking self-contained intent; (v) discard prompts without a clear task intention; and (vi) validate PII triggers with a GPT-4o checker, confirming at least three *unique*, text-based entities per message. For WildChat, we additionally restrict to messages without pre-existing redactions. The final sets comprise over 500 ShareGPT messages and over 100 WildChat messages. The ShareGPT pool supports replacement-pair creation for the human study and comparator distillation, and both sets are later reused for pipeline operationalization.

D DETECTION AND ABSTRACTION SYSTEM PROMPT

D.1 DETECTION

You are an expert in cybersecurity and data privacy. You are now tasked to detect PII from the given text, using the following taxonomy only:

- `NAME`: Name
- `EMAIL`: Email address
- `PHONE_NUMBER`: Phone number
- `ID`: Identifiers, including ID Number, passport number, SSN, driver's license, taxpayer, identification number
- `ONLINE_IDENTITY`: IP address, username, URL, password, key
- `GEOLOCATION`: Places and locations and addresses, such as cities, provinces, countries, international regions, or named infrastructures (bus stops, bridges, etc.).
- `AFFILIATION`: Names of organizations, such as public and private companies, schools, universities, public institutions, prisons, healthcare institutions, non-governmental organizations, churches, etc.
- `DEMOGRAPHIC_ATTRIBUTE`: Demographic attributes of a person, such as native language, descent, heritage, ethnicity, nationality, religious or political group, birthmarks, ages, sexual orientation, gender and sex.
- `TIME`: Description of a specific date, time, or duration.
- `HEALTH_INFORMATION`: Details concerning an individual's health status, medical conditions, treatment records, and health insurance information.
- `FINANCIAL_INFORMATION`: Financial details such as bank account numbers, credit card numbers, investment records, salary information, and other financial statuses or activities.

- `EDUCATIONAL_RECORD`: Educational background details, including academic records, transcripts, degrees, and certification.

For the given message that a user sends to a chatbot, identify all the personally identifiable information using the above taxonomy only, and the `entity_type` should be selected from the all-caps categories. Note that the information should be related to a real person not in a public context, but okay if not uniquely identifiable. Result should be in its minimum possible unit.

Return *only* a JSON object in the following format: `"results": ["entity_type": YOU_DECIDE_THE_PII_TYPE, "text": PART_OF_MESSAGE_YOU_IDENTIFIED_AS_PII]`

D.2 Abstraction

> Rewrite the text to abstract the protected information, without changing other parts. Please output in JSON format only.

Input:

```
<Text>I graduated from CMU, and I earn a six-figure salary.
Today a lady from China said...</Text>
<ProtectedInformation1>CMU</ProtectedInformation1>
<ProtectedInformation2>China</ProtectedInformation2>
```

Output:

```
{"results": [{"protected": "CMU", "abstracted": "a prestigious
American university"},
{"protected": "China", "abstracted": "A country in Asia"}]}
```

E Utility Check

Utility Evaluator System Prompt You are a utility evaluator. Your task is to assess whether a second response adequately addresses a user's original message. You will be provided with: (i) the user message (what the user is asking or talking about), (ii) a reference response A (a standard answer that is known to be good), and (iii) a candidate response B (a second version that needs to be evaluated). You must decide whether the candidate response is "PASS" or "FAIL". This is not a test of exact wording or stylistic similarity; focus on whether the candidate response addresses all key points or needs expressed in the user message. If it does, it passes. If it fails to address one or more key points, it fails.

Input will contain three tagged sections: <user_message>, <response_A>, and <response_B>.

Return JSON with the result and a one-line explanation: `"Result": "PASS/FAIL", "Reason": YOUR_EXPLANATION`

F Utility Threshold Sensitivity Study

To evaluate how sensitive our method is to the choice of the utility threshold γ (§3.1) in the open-ended task setting, we performed an additional experiment on the ShareGPT and WildChat datasets. In the main paper, the utility predicate is binary because the evaluator (GPT-4o) returns categorical judgments. For this study, we modified the evaluator's instruction so that it produces a numerical utility score between 1 and 10, where the model's original response is defined as a score of 10. A candidate transformation passes the utility predicate if its score is at least $10 - \gamma$.

We evaluated three thresholds, $\gamma = 0, 1, 2$, corresponding to minimum required utility scores of 10, 9, and 8. All other components of the pipeline remained unchanged.

To determine whether the resulting quality differences are perceptible to users, we conducted a human evaluation using Prolific. We recruited 90 English speaking participants located in the United States. For each of 90 sampled prompts, we generated pairwise comparisons between (i) the original response and (ii) a response produced under one of the γ settings. Participants selected whether A, B, or SAME was of higher quality.

The results were as follows:

- $\gamma = 0$: SAME = 36.8%, original preferred = 32.6%, masked preferred = 30.5%
- $\gamma = 1$: SAME = 30.9%, original preferred = 42.5%, masked preferred = 26.6%
- $\gamma = 2$: SAME = 28.1%, original preferred = 44.1%, masked preferred = 27.8%

The trend is clear. As γ increases, the proportion of SAME judgments decreases and preference for the original output increases. This shows that even small relaxations of the utility threshold produce user perceptible quality degradation. These findings justify the use of a strict binary predicate 3.2, which prohibits any reduction in utility and therefore represents a conservative and safety oriented choice for minimizing disclosure.

G PROMPT SPECIFICATIONS FOR ZERO-SHOT PREDICTION

G.1 SMALL-MODEL FAILURE.

We mark the smallest-capacity model (Qwen2.5–0.5B) as N/A for the single-pass prediction due to persistent schema/IO failures: (i) emitting actions outside {RETAIN, ABSTRACT, REDACT}; (ii) using PII *types* as keys instead of entity IDs; and (iii) remaining unstable even with reduced prompts. Notably, when we prefill an all-RETAIN map *inside the prompt* to constrain decoding, the model either (a) echoes the seed map (no decision-making) verbatim or (b) still produces noncompliant/irrelevant text, failing to follow the output schema. Given this degenerate behavior, a single-pass prediction is not meaningful so we did not use this model as predictor in the experiments.

G.2 SYSTEM PROMPTS

Open-ended tasks (ShareGPT / WildChat). You are a PII Minimization Transformer. Return *one line*: a single JSON object (no markdown/prose). **Goal:** for each PII surface string in pii_dict, choose the most privacy-preserving action that keeps response quality for open-ended tasks. **Order:** RETAIN $<$ ABSTRACT $<$ REDACT (prefer stronger only if quality is unchanged). **Schema:** "transformation":"<PII_KEY>":"retain|abstract|redact", ..., "explanation":"ONE SHORT SENTENCE, NO RAW PII".

Rules: (1) Keys must be the exact surface strings from pii_dict (verbatim; do not add/remove/rename). (2) Values are exactly one of retain|abstract|redact (lowercase). (3) If an action is chosen, it applies to all variants in variants_map. (4) If the needed entry is missing in abstract_map/redact_map, default to retain for that key. (5) Keep key order aligned with pii_dict if possible.

Deterministic tasks (MedQA / CaseHOLD). Same as (A) except the **Goal**: choose actions that still allow selecting the correct option in a multiple-choice setting. Use the same order, schema, and rules.

Tiny-model strict prompt (values-only edits). Return *one line* JSON only: {"transformation":{...}}. Inputs: pii_dict and draft_transformation where every key (verbatim from pii_dict) is preset to "retain". **Edit only the values** to one of retain|abstract|redact. Do not add/remove/rename any key.

G.3 USER PAYLOADS

Regular user payload (with maps). Fields: message (original text), pii_dict ({surface: type}), variants_map ({surface: [aliases]}), redact_map/abstract_map (per-surface replacements).

"Edit values only to one of retain|abstract|redact; do not change keys. Return {"transformation": <edited draft>}."

G.4 Ablation of Minimization-Order

To test whether the observed preference for ABSTRACT arises from our prompt design, we conducted two ablations of the system instruction used for single-pass prediction. The first variant (*order_only*) removes the clause "prefer stronger only if quality is unchanged," while the second (*no order/notion*) removes the entire minimization-order line. We ran GPT–5, Mistral–24B, and Qwen2.5–7B on ShareGPT and MedQA using all three prompt types, keeping all other settings fixed. These three models were chosen to cover a representative range of capacities and training regimes, and ShareGPT (open-ended) and MedQA (closed-ended) serve as one exemplar dataset for each task type. As shown in Table 4, removing these instructions does not eliminate the strong preference for ABSTRACT; redaction increases marginally, but the overall pattern remains unchanged. This indicates that the abstraction bias is not prompt-induced but reflects a model-internal tendency toward fluency-preserving transformations.

dataset	model	prompt_type	redact	abstract	retain	undecided
ShareGPT	gpt-5	order+notion	164 (13.8%)	663 (55.7%)	363 (30.5%)	0 (0.0%)
ShareGPT	gpt-5	order_only	225 (18.9%)	656 (55.1%)	309 (26.0%)	0 (0.0%)
ShareGPT	gpt-5	no order/notion	153 (12.9%)	728 (61.2%)	309 (26.0%)	0 (0.0%)
ShareGPT	mistral-small-24b	order+notion	6 (0.5%)	1155 (97.1%)	29 (2.4%)	0 (0.0%)
ShareGPT	mistral-small-24b	order_only	20 (1.7%)	1070 (89.9%)	100 (8.4%)	0 (0.0%)
ShareGPT	mistral-small-24b	no order/notion	20 (1.7%)	1027 (86.3%)	143 (12.0%)	0 (0.0%)
ShareGPT	qwen2.5-7b	order+notion	16 (1.3%)	1174 (98.7%)	0 (0.0%)	0 (0.0%)
ShareGPT	qwen2.5-7b	order_only	38 (3.2%)	1116 (93.8%)	31 (2.6%)	5 (0.4%)
ShareGPT	qwen2.5-7b	no order/notion	139 (11.7%)	1000 (84.0%)	46 (3.9%)	5 (0.4%)
MedQA	gpt-5	order+notion	533 (54.6%)	405 (41.5%)	38 (3.9%)	0 (0.0%)
MedQA	gpt-5	order_only	808 (82.8%)	140 (14.3%)	28 (2.9%)	0 (0.0%)
MedQA	gpt-5	no order/notion	688 (70.5%)	258 (26.4%)	30 (3.1%)	0 (0.0%)
MedQA	mistral-small-24b	order+notion	0 (0.0%)	912 (93.4%)	64 (6.6%)	0 (0.0%)
MedQA	mistral-small-24b	order_only	2 (0.2%)	609 (62.4%)	365 (37.4%)	0 (0.0%)
MedQA	mistral-small-24b	no order/notion	2 (0.2%)	531 (54.4%)	443 (45.4%)	0 (0.0%)
MedQA	qwen2.5-7b	order+notion	0 (0.0%)	973 (99.7%)	3 (0.3%)	0 (0.0%)
MedQA	qwen2.5-7b	order_only	2 (0.2%)	940 (96.3%)	30 (3.1%)	4 (0.4%)
MedQA	qwen2.5-7b	no order/notion	19 (1.9%)	928 (95.1%)	28 (2.9%)	1 (0.1%)

Table 4: Ablation of minimization-order instructions across models and datasets (without total-pii column).

H Pipeline + self prediction

H.1 CASEHOLD

Type	**Weighted Redact**	**Weighted Abstract**	**Weighted Retain**
NAME	98.94% (93/94)	0.00% (0/94)	1.06% (1/94)
TIME	96.61% (57/59)	1.69% (1/59)	1.69% (1/59)
GEOLOCATION	96.15% (75/78)	1.28% (1/78)	2.56% (2/78)
AFFILIATION	93.17% (150/161)	1.86% (3/161)	4.97% (8/161)
RACE	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
AGE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
HEALTH_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	95.77% (385/402)	1.24% (5/402)	2.99% (12/402)

Table 5: Weighted Results per Type and Overall (Oracle for **CaseHOLD**), Model: **gpt-4.1-nano**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	98.94% (93/94)	0.00% (0/94)	1.06% (1/94)
TIME	100.00% (59/59)	0.00% (0/59)	0.00% (0/59)
GEOLOCATION	100.00% (78/78)	0.00% (0/78)	0.00% (0/78)
AFFILIATION	100.00% (161/161)	0.00% (0/161)	0.00% (0/161)
RACE	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
AGE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
HEALTH_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	99.75% (401/402)	0.00% (0/402)	0.25% (1/402)

Table 6: Weighted Results per Type and Overall (Oracle for **CaseHOLD**), Model: **gpt-4.1**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	97.87% (92/94)	2.13% (2/94)	0.00% (0/94)
TIME	96.61% (57/59)	1.69% (1/59)	1.69% (1/59)
GEOLOCATION	98.72% (77/78)	1.28% (1/78)	0.00% (0/78)
AFFILIATION	100.00% (161/161)	0.00% (0/161)	0.00% (0/161)
RACE	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
AGE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
HEALTH_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	98.76% (397/402)	1.00% (4/402)	0.25% (1/402)

Table 7: Weighted Results per Type and Overall (Oracle for **CaseHOLD**), Model: **gpt-5**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	100.00% (94/94)	0.00% (0/94)	0.00% (0/94)
TIME	98.31% (58/59)	1.69% (1/59)	0.00% (0/59)
GEOLOCATION	100.00% (78/78)	0.00% (0/78)	0.00% (0/78)
AFFILIATION	99.38% (160/161)	0.62% (1/161)	0.00% (0/161)
RACE	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
AGE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
HEALTH_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	99.50% (400/402)	0.50% (2/402)	0.00% (0/402)

Table 8: Weighted Results per Type and Overall (Oracle for **CaseHOLD**), Model: **claude-3-7-sonnet-20250219**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	100.00% (94/94)	0.00% (0/94)	0.00% (0/94)
TIME	100.00% (59/59)	0.00% (0/59)	0.00% (0/59)
GEOLOCATION	100.00% (78/78)	0.00% (0/78)	0.00% (0/78)
AFFILIATION	100.00% (161/161)	0.00% (0/161)	0.00% (0/161)
RACE	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
AGE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
HEALTH_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	100.00% (402/402)	0.00% (0/402)	0.00% (0/402)

Table 9: Weighted Results per Type and Overall (Oracle for **CaseHOLD**), Model: **claude-sonnet-4-20250514**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	84.04% (79/94)	5.32% (5/94)	10.64% (10/94)
TIME	57.63% (34/59)	6.78% (4/59)	35.59% (21/59)
GEOLOCATION	74.36% (58/78)	5.13% (4/78)	20.51% (16/78)
AFFILIATION	71.43% (115/161)	9.94% (16/161)	18.63% (30/161)
RACE	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
AGE	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
HEALTH_INFORMATION	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
FINANCIAL_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	73.13% (294/402)	7.21% (29/402)	19.65% (79/402)

Table 10: Weighted Results per Type and Overall (Oracle for **CaseHOLD**), Model: **lgai/exaone-deep-32b**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	95.74% (90/94)	0.00% (0/94)	4.26% (4/94)
TIME	96.61% (57/59)	1.69% (1/59)	1.69% (1/59)
GEOLOCATION	92.31% (72/78)	1.28% (1/78)	6.41% (5/78)
AFFILIATION	95.65% (154/161)	0.62% (1/161)	3.73% (6/161)
RACE	75.00% (3/4)	25.00% (1/4)	0.00% (0/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
AGE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
HEALTH_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	95.02% (382/402)	1.00% (4/402)	3.98% (16/402)

Table 11: Weighted Results per Type and Overall (Oracle for **CaseHOLD**), Model: **mistralai/mistral-small-3.1-24b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	95.74% (90/94)	1.06% (1/94)	3.19% (3/94)
TIME	98.31% (58/59)	0.00% (0/59)	1.69% (1/59)
GEOLOCATION	97.44% (76/78)	0.00% (0/78)	2.56% (2/78)
AFFILIATION	97.52% (157/161)	0.00% (0/161)	2.48% (4/161)
RACE	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
AGE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
HEALTH_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	97.01% (390/402)	0.25% (1/402)	2.74% (11/402)

Table 12: Weighted Results per Type and Overall (Oracle for **CaseHOLD**), Model: **qwen/qwen2.5-7b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	39.36% (37/94)	5.32% (5/94)	55.32% (52/94)
TIME	35.59% (21/59)	5.08% (3/59)	59.32% (35/59)
GEOLOCATION	38.46% (30/78)	8.97% (7/78)	52.56% (41/78)
AFFILIATION	44.10% (71/161)	6.21% (10/161)	49.69% (80/161)
RACE	25.00% (1/4)	25.00% (1/4)	50.00% (2/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
INCOME	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
AGE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
HEALTH_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	40.80% (164/402)	6.47% (26/402)	52.74% (212/402)

Table 13: Weighted Results per Type and Overall (Oracle for **CaseHOLD**), Model: **qwen/qwen2.5-0.5b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	71.28% (67/94)	28.72% (27/94)	0.00% (0/94)
TIME	79.66% (47/59)	20.34% (12/59)	0.00% (0/59)
GEOLOCATION	80.77% (63/78)	19.23% (15/78)	0.00% (0/78)
AFFILIATION	81.99% (132/161)	18.01% (29/161)	0.00% (0/161)
RACE	75.00% (3/4)	25.00% (1/4)	0.00% (0/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
AGE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
HEALTH_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	78.36% (315/402)	21.64% (87/402)	0.00% (0/402)

Table 14: Weighted Results per Type and Overall (Prediction for **CaseHOLD**), Model: **gpt-4.1-nano**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	0.00% (0/94)	100.00% (94/94)	0.00% (0/94)
TIME	0.00% (0/59)	100.00% (59/59)	0.00% (0/59)
GEOLOCATION	0.00% (0/78)	84.62% (66/78)	15.38% (12/78)
AFFILIATION	0.00% (0/161)	95.03% (153/161)	4.97% (8/161)
RACE	0.00% (0/4)	75.00% (3/4)	25.00% (1/4)
ETHNICITY	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
AGE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
HEALTH_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	0.00% (0/402)	94.53% (380/402)	5.47% (22/402)

Table 15: Weighted Results per Type and Overall (Prediction for **CaseHOLD**), Model: **gpt-4.1**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	88.30% (83/94)	6.38% (6/94)	5.32% (5/94)
TIME	86.44% (51/59)	13.56% (8/59)	0.00% (0/59)
GEOLOCATION	76.92% (60/78)	10.26% (8/78)	12.82% (10/78)
AFFILIATION	89.44% (144/161)	5.59% (9/161)	4.97% (8/161)
RACE	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
AGE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
HEALTH_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	86.57% (348/402)	7.71% (31/402)	5.72% (23/402)

Table 16: Weighted Results per Type and Overall (Prediction for **CaseHOLD**), Model: **gpt-5**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	1.06% (1/94)	62.77% (59/94)	36.17% (34/94)
TIME	1.69% (1/59)	45.76% (27/59)	52.54% (31/59)
GEOLOCATION	0.00% (0/78)	20.51% (16/78)	79.49% (62/78)
AFFILIATION	0.00% (0/161)	32.30% (52/161)	67.70% (109/161)
RACE	0.00% (0/4)	50.00% (2/4)	50.00% (2/4)
ETHNICITY	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
AGE	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
HEALTH_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	0.50% (2/402)	39.80% (160/402)	59.70% (240/402)

Table 17: Weighted Results per Type and Overall (Prediction for **CaseHOLD**), Model: **claude-3-7-sonnet-20250219**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	17.02% (16/94)	60.64% (57/94)	22.34% (21/94)
TIME	11.86% (7/59)	62.71% (37/59)	25.42% (15/59)
GEOLOCATION	0.00% (0/78)	29.49% (23/78)	70.51% (55/78)
AFFILIATION	3.11% (5/161)	35.40% (57/161)	61.49% (99/161)
RACE	0.00% (0/4)	50.00% (2/4)	50.00% (2/4)
ETHNICITY	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
AGE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
HEALTH_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	6.97% (28/402)	44.78% (180/402)	48.26% (194/402)

Table 18: Weighted Results per Type and Overall (Prediction for **CaseHOLD**), Model: **claude-sonnet-4-20250514**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	70.21% (66/94)	29.79% (28/94)	0.00% (0/94)
TIME	66.10% (39/59)	33.90% (20/59)	0.00% (0/59)
GEOLOCATION	66.67% (52/78)	33.33% (26/78)	0.00% (0/78)
AFFILIATION	72.05% (116/161)	27.95% (45/161)	0.00% (0/161)
RACE	75.00% (3/4)	25.00% (1/4)	0.00% (0/4)
ETHNICITY	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
AGE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
HEALTH_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	69.90% (281/402)	30.10% (121/402)	0.00% (0/402)

Table 19: Weighted Results per Type and Overall (Prediction for **CaseHOLD**), Model: **lgai/exaone-deep-32b**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	0.00% (0/94)	100.00% (94/94)	0.00% (0/94)
TIME	0.00% (0/59)	100.00% (59/59)	0.00% (0/59)
GEOLOCATION	1.28% (1/78)	98.72% (77/78)	0.00% (0/78)
AFFILIATION	0.00% (0/161)	99.38% (160/161)	0.62% (1/161)
RACE	0.00% (0/4)	100.00% (4/4)	0.00% (0/4)
ETHNICITY	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
AGE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
HEALTH_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	0.25% (1/402)	99.50% (400/402)	0.25% (1/402)

Table 20: Weighted Results per Type and Overall (Prediction for **CaseHOLD**), Model: **mistralai/mistral-small-3.1-24b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	3.19% (3/94)	96.81% (91/94)	0.00% (0/94)
TIME	5.08% (3/59)	94.92% (56/59)	0.00% (0/59)
GEOLOCATION	1.28% (1/78)	98.72% (77/78)	0.00% (0/78)
AFFILIATION	4.97% (8/161)	95.03% (153/161)	0.00% (0/161)
RACE	25.00% (1/4)	75.00% (3/4)	0.00% (0/4)
ETHNICITY	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
AGE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
HEALTH_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	3.98% (16/402)	96.02% (386/402)	0.00% (0/402)

Table 21: Weighted Results per Type and Overall (Prediction for **CaseHOLD**), Model: **qwen/qwen2.5-7b-instruct**

H.2 MEDQA

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	96.46% (109/113)	0.00% (0/113)	3.54% (4/113)
GENDER	98.28% (57/58)	0.00% (0/58)	1.72% (1/58)
OCCUPATION	100.00% (9/9)	0.00% (0/9)	0.00% (0/9)
HEALTH_INFORMATION	87.08% (647/743)	2.96% (22/743)	9.96% (74/743)
GEOLOCATION	94.74% (18/19)	0.00% (0/19)	5.26% (1/19)
RACE	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	95.00% (19/20)	0.00% (0/20)	5.00% (1/20)
SEXUAL_ORIENTATION	100.00% (3/3)	0.00% (0/3)	0.00% (0/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	89.45% (873/976)	2.25% (22/976)	8.30% (81/976)

Table 22: Weighted Results per Type and Overall (Oracle for **MedQA**), Model: **gpt-4.1-nano**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	98.23% (111/113)	1.77% (2/113)	0.00% (0/113)
GENDER	96.55% (56/58)	1.72% (1/58)	1.72% (1/58)
OCCUPATION	100.00% (9/9)	0.00% (0/9)	0.00% (0/9)
HEALTH_INFORMATION	96.90% (720/743)	1.48% (11/743)	1.62% (12/743)
GEOLOCATION	100.00% (19/19)	0.00% (0/19)	0.00% (0/19)
RACE	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	100.00% (20/20)	0.00% (0/20)	0.00% (0/20)
SEXUAL_ORIENTATION	100.00% (3/3)	0.00% (0/3)	0.00% (0/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	97.23% (949/976)	1.43% (14/976)	1.33% (13/976)

Table 23: Weighted Results per Type and Overall (Oracle for **MedQA**), Model: **gpt-4.1**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	100.00% (113/113)	0.00% (0/113)	0.00% (0/113)
GENDER	100.00% (58/58)	0.00% (0/58)	0.00% (0/58)
OCCUPATION	100.00% (9/9)	0.00% (0/9)	0.00% (0/9)
HEALTH_INFORMATION	95.42% (709/743)	2.69% (20/743)	1.88% (14/743)
GEOLOCATION	94.74% (18/19)	5.26% (1/19)	0.00% (0/19)
RACE	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	100.00% (20/20)	0.00% (0/20)	0.00% (0/20)
SEXUAL_ORIENTATION	100.00% (3/3)	0.00% (0/3)	0.00% (0/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	96.41% (941/976)	2.15% (21/976)	1.43% (14/976)

Table 24: Weighted Results per Type and Overall (Oracle for **MedQA**), Model: **gpt-5**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	79.65% (90/113)	10.62% (12/113)	9.73% (11/113)
GENDER	70.69% (41/58)	13.79% (8/58)	15.52% (9/58)
OCCUPATION	88.89% (8/9)	11.11% (1/9)	0.00% (0/9)
HEALTH_INFORMATION	69.99% (520/743)	14.27% (106/743)	15.75% (117/743)
GEOLOCATION	78.95% (15/19)	21.05% (4/19)	0.00% (0/19)
RACE	87.50% (7/8)	0.00% (0/8)	12.50% (1/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	55.00% (11/20)	25.00% (5/20)	20.00% (4/20)
SEXUAL_ORIENTATION	33.33% (1/3)	33.33% (1/3)	33.33% (1/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	71.31% (696/976)	14.04% (137/976)	14.65% (143/976)

Table 25: Weighted Results per Type and Overall (Oracle for **MedQA**), Model: **claude-3-7-sonnet-20250219**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	100.00% (113/113)	0.00% (0/113)	0.00% (0/113)
GENDER	100.00% (58/58)	0.00% (0/58)	0.00% (0/58)
OCCUPATION	100.00% (9/9)	0.00% (0/9)	0.00% (0/9)
HEALTH_INFORMATION	94.75% (704/743)	3.50% (26/743)	1.75% (13/743)
GEOLOCATION	100.00% (19/19)	0.00% (0/19)	0.00% (0/19)
RACE	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	100.00% (20/20)	0.00% (0/20)	0.00% (0/20)
SEXUAL_ORIENTATION	100.00% (3/3)	0.00% (0/3)	0.00% (0/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	96.00% (937/976)	2.66% (26/976)	1.33% (13/976)

Table 26: Weighted Results per Type and Overall (Oracle for **MedQA**), Model: **claude-sonnet-4-20250514**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	80.53% (91/113)	11.50% (13/113)	7.96% (9/113)
GENDER	82.76% (48/58)	6.90% (4/58)	10.34% (6/58)
OCCUPATION	77.78% (7/9)	22.22% (2/9)	0.00% (0/9)
HEALTH_INFORMATION	74.16% (551/743)	11.98% (89/743)	13.86% (103/743)
GEOLOCATION	84.21% (16/19)	5.26% (1/19)	10.53% (2/19)
RACE	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	70.00% (14/20)	15.00% (3/20)	15.00% (3/20)
SEXUAL_ORIENTATION	66.67% (2/3)	0.00% (0/3)	33.33% (1/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	75.82% (740/976)	11.48% (112/976)	12.70% (124/976)

Table 27: Weighted Results per Type and Overall (Oracle for **MedQA**), Model: **lgai/exaone-deep-32b**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	100.00% (113/113)	0.00% (0/113)	0.00% (0/113)
GENDER	100.00% (58/58)	0.00% (0/58)	0.00% (0/58)
OCCUPATION	88.89% (8/9)	11.11% (1/9)	0.00% (0/9)
HEALTH_INFORMATION	96.37% (716/743)	2.29% (17/743)	1.35% (10/743)
GEOLOCATION	100.00% (19/19)	0.00% (0/19)	0.00% (0/19)
RACE	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	100.00% (20/20)	0.00% (0/20)	0.00% (0/20)
SEXUAL_ORIENTATION	100.00% (3/3)	0.00% (0/3)	0.00% (0/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	97.03% (947/976)	1.95% (19/976)	1.02% (10/976)

Table 28: Weighted Results per Type and Overall (Oracle for **MedQA**), Model: **mistralai/mistral-small-3.1-24b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	94.69% (107/113)	2.65% (3/113)	2.65% (3/113)
GENDER	93.10% (54/58)	5.17% (3/58)	1.72% (1/58)
OCCUPATION	88.89% (8/9)	11.11% (1/9)	0.00% (0/9)
HEALTH_INFORMATION	88.29% (656/743)	6.86% (51/743)	4.85% (36/743)
GEOLOCATION	89.47% (17/19)	10.53% (2/19)	0.00% (0/19)
RACE	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	90.00% (18/20)	10.00% (2/20)	0.00% (0/20)
SEXUAL_ORIENTATION	100.00% (3/3)	0.00% (0/3)	0.00% (0/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	89.45% (873/976)	6.45% (63/976)	4.10% (40/976)

Table 29: Weighted Results per Type and Overall (Oracle for **MedQA**), Model: **qwen/qwen2.5-7b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	42.48% (48/113)	13.27% (15/113)	44.25% (50/113)
GENDER	39.66% (23/58)	10.34% (6/58)	50.00% (29/58)
OCCUPATION	55.56% (5/9)	11.11% (1/9)	33.33% (3/9)
HEALTH_INFORMATION	24.09% (179/743)	14.27% (106/743)	61.64% (458/743)
GEOLOCATION	31.58% (6/19)	15.79% (3/19)	52.63% (10/19)
RACE	50.00% (4/8)	12.50% (1/8)	37.50% (3/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	45.00% (9/20)	10.00% (2/20)	45.00% (9/20)
SEXUAL_ORIENTATION	33.33% (1/3)	33.33% (1/3)	33.33% (1/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	28.48% (278/976)	13.83% (135/976)	57.68% (563/976)

Table 30: Weighted Results per Type and Overall (Oracle for **MedQA**), Model: **qwen/qwen2.5-0.5b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	41.59% (47/113)	58.41% (66/113)	0.00% (0/113)
GENDER	37.93% (22/58)	60.34% (35/58)	1.72% (1/58)
OCCUPATION	66.67% (6/9)	33.33% (3/9)	0.00% (0/9)
HEALTH_INFORMATION	21.27% (158/743)	60.97% (453/743)	17.77% (132/743)
GEOLOCATION	57.89% (11/19)	42.11% (8/19)	0.00% (0/19)
RACE	37.50% (3/8)	62.50% (5/8)	0.00% (0/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	65.00% (13/20)	35.00% (7/20)	0.00% (0/20)
SEXUAL_ORIENTATION	33.33% (1/3)	66.67% (2/3)	0.00% (0/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	27.05% (264/976)	59.32% (579/976)	13.63% (133/976)

Table 31: Weighted Results per Type and Overall (Prediction for **MedQA**), Model: **gpt-4.1-nano**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	0.00% (0/113)	99.12% (112/113)	0.88% (1/113)
GENDER	0.00% (0/58)	98.28% (57/58)	1.72% (1/58)
OCCUPATION	0.00% (0/9)	100.00% (9/9)	0.00% (0/9)
HEALTH_INFORMATION	0.00% (0/743)	98.52% (732/743)	1.48% (11/743)
GEOLOCATION	10.53% (2/19)	89.47% (17/19)	0.00% (0/19)
RACE	0.00% (0/8)	100.00% (8/8)	0.00% (0/8)
MARITAL_STATUS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
TIME	0.00% (0/20)	100.00% (20/20)	0.00% (0/20)
SEXUAL_ORIENTATION	0.00% (0/3)	100.00% (3/3)	0.00% (0/3)
AFFILIATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
DIETARY_PREFERENCE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	0.20% (2/976)	98.46% (961/976)	1.33% (13/976)

Table 32: Weighted Results per Type and Overall (Prediction for **MedQA**), Model: **gpt-4.1**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	57.52% (65/113)	41.59% (47/113)	0.88% (1/113)
GENDER	74.14% (43/58)	22.41% (13/58)	3.45% (2/58)
OCCUPATION	66.67% (6/9)	33.33% (3/9)	0.00% (0/9)
HEALTH_INFORMATION	51.82% (385/743)	43.47% (323/743)	4.71% (35/743)
GEOLOCATION	57.89% (11/19)	42.11% (8/19)	0.00% (0/19)
RACE	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	50.00% (10/20)	50.00% (10/20)	0.00% (0/20)
SEXUAL_ORIENTATION	100.00% (3/3)	0.00% (0/3)	0.00% (0/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	54.61% (533/976)	41.50% (405/976)	3.89% (38/976)

Table 33: Weighted Results per Type and Overall (Prediction for **MedQA**), Model: **gpt-5**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	0.00% (0/113)	30.09% (34/113)	69.91% (79/113)
GENDER	0.00% (0/58)	22.41% (13/58)	77.59% (45/58)
OCCUPATION	0.00% (0/9)	55.56% (5/9)	44.44% (4/9)
HEALTH_INFORMATION	0.00% (0/743)	2.96% (22/743)	97.04% (721/743)
GEOLOCATION	0.00% (0/19)	52.63% (10/19)	47.37% (9/19)
RACE	25.00% (2/8)	62.50% (5/8)	12.50% (1/8)
MARITAL_STATUS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
TIME	0.00% (0/20)	20.00% (4/20)	80.00% (16/20)
SEXUAL_ORIENTATION	0.00% (0/3)	66.67% (2/3)	33.33% (1/3)
AFFILIATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
DIETARY_PREFERENCE	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
Overall	0.20% (2/976)	9.94% (97/976)	89.86% (877/976)

Table 34: Weighted Results per Type and Overall (Prediction for **MedQA**), Model: **claude-3-7-sonnet-20250219**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	0.00% (0/113)	70.80% (80/113)	29.20% (33/113)
GENDER	1.72% (1/58)	74.14% (43/58)	24.14% (14/58)
OCCUPATION	0.00% (0/9)	100.00% (9/9)	0.00% (0/9)
HEALTH_INFORMATION	0.00% (0/743)	15.21% (113/743)	84.79% (630/743)
GEOLOCATION	5.26% (1/19)	63.16% (12/19)	31.58% (6/19)
RACE	87.50% (7/8)	0.00% (0/8)	12.50% (1/8)
MARITAL_STATUS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
TIME	0.00% (0/20)	40.00% (8/20)	60.00% (12/20)
SEXUAL_ORIENTATION	0.00% (0/3)	100.00% (3/3)	0.00% (0/3)
AFFILIATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
DIETARY_PREFERENCE	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
Overall	0.92% (9/976)	27.66% (270/976)	71.41% (697/976)

Table 35: Weighted Results per Type and Overall (Prediction for **MedQA**), Model: **claude-sonnet-4-20250514**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	84.07% (95/113)	15.93% (18/113)	0.00% (0/113)
GENDER	82.76% (48/58)	15.52% (9/58)	1.72% (1/58)
OCCUPATION	88.89% (8/9)	11.11% (1/9)	0.00% (0/9)
HEALTH_INFORMATION	45.76% (340/743)	54.10% (402/743)	0.13% (1/743)
GEOLOCATION	94.74% (18/19)	5.26% (1/19)	0.00% (0/19)
RACE	87.50% (7/8)	12.50% (1/8)	0.00% (0/8)
MARITAL_STATUS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
TIME	80.00% (16/20)	20.00% (4/20)	0.00% (0/20)
SEXUAL_ORIENTATION	66.67% (2/3)	33.33% (1/3)	0.00% (0/3)
AFFILIATION	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
DIETARY_PREFERENCE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	54.92% (536/976)	44.88% (438/976)	0.20% (2/976)

Table 36: Weighted Results per Type and Overall (Prediction for **MedQA**), Model: **lgai/exaone-deep-32b**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	0.00% (0/113)	100.00% (113/113)	0.00% (0/113)
GENDER	0.00% (0/58)	100.00% (58/58)	0.00% (0/58)
OCCUPATION	0.00% (0/9)	100.00% (9/9)	0.00% (0/9)
HEALTH_INFORMATION	0.00% (0/743)	91.39% (679/743)	8.61% (64/743)
GEOLOCATION	0.00% (0/19)	100.00% (19/19)	0.00% (0/19)
RACE	0.00% (0/8)	100.00% (8/8)	0.00% (0/8)
MARITAL_STATUS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
TIME	0.00% (0/20)	100.00% (20/20)	0.00% (0/20)
SEXUAL_ORIENTATION	0.00% (0/3)	100.00% (3/3)	0.00% (0/3)
AFFILIATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
DIETARY_PREFERENCE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	0.00% (0/976)	93.44% (912/976)	6.56% (64/976)

Table 37: Weighted Results per Type and Overall (Prediction for **MedQA**), Model: **mistralai/mistral-small-3.1-24b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
AGE	0.00% (0/113)	100.00% (113/113)	0.00% (0/113)
GENDER	0.00% (0/58)	96.55% (56/58)	3.45% (2/58)
OCCUPATION	0.00% (0/9)	100.00% (9/9)	0.00% (0/9)
HEALTH_INFORMATION	0.00% (0/743)	99.87% (742/743)	0.13% (1/743)
GEOLOCATION	0.00% (0/19)	100.00% (19/19)	0.00% (0/19)
RACE	0.00% (0/8)	100.00% (8/8)	0.00% (0/8)
MARITAL_STATUS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
TIME	0.00% (0/20)	100.00% (20/20)	0.00% (0/20)
SEXUAL_ORIENTATION	0.00% (0/3)	100.00% (3/3)	0.00% (0/3)
AFFILIATION	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
DIETARY_PREFERENCE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	0.00% (0/976)	99.69% (973/976)	0.31% (3/976)

Table 38: Weighted Results per Type and Overall (Prediction for **MedQA**), Model: **qwen/qwen2.5-7b-instruct**

H.3 SHAREGPT

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	91.28% (157/172)	5.23% (9/172)	3.49% (6/172)
AFFILIATION	90.64% (155/171)	5.26% (9/171)	4.09% (7/171)
TIME	67.05% (177/264)	7.58% (20/264)	25.38% (67/264)
URL	75.00% (15/20)	20.00% (4/20)	5.00% (1/20)
EMAIL	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
GEOLOCATION	66.67% (218/327)	17.13% (56/327)	16.21% (53/327)
RELIGION	0.00% (0/2)	50.00% (1/2)	50.00% (1/2)
FINANCIAL_INFORMATION	72.73% (8/11)	18.18% (2/11)	9.09% (1/11)
MARITAL_STATUS	63.64% (7/11)	27.27% (3/11)	9.09% (1/11)
OCCUPATION	83.33% (50/60)	11.67% (7/60)	5.00% (3/60)
VEHICLE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	87.50% (7/8)	12.50% (1/8)	0.00% (0/8)
HEALTH_INFORMATION	65.31% (32/49)	16.33% (8/49)	18.37% (9/49)
EDUCATIONAL_RECORD	100.00% (10/10)	0.00% (0/10)	0.00% (0/10)
AGE	67.86% (38/56)	26.79% (15/56)	5.36% (3/56)
GENDER	61.54% (8/13)	23.08% (3/13)	15.38% (2/13)
ETHNICITY	80.00% (4/5)	20.00% (1/5)	0.00% (0/5)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
IP_ADDRESS	66.67% (2/3)	0.00% (0/3)	33.33% (1/3)
RACE	50.00% (1/2)	50.00% (1/2)	0.00% (0/2)
Overall	75.04% (893/1190)	11.93% (142/1190)	13.03% (155/1190)

Table 39: Weighted Results per Type and Overall (Oracle for **ShareGPT90K**), Model: **gpt-4.1-nano**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	87.79% (151/172)	6.98% (12/172)	5.23% (9/172)
AFFILIATION	96.49% (165/171)	1.75% (3/171)	1.75% (3/171)
TIME	78.03% (206/264)	10.61% (28/264)	11.36% (30/264)
URL	85.00% (17/20)	15.00% (3/20)	0.00% (0/20)
EMAIL	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
GEOLOCATION	66.97% (219/327)	22.32% (73/327)	10.70% (35/327)
RELIGION	50.00% (1/2)	50.00% (1/2)	0.00% (0/2)
FINANCIAL_INFORMATION	81.82% (9/11)	18.18% (2/11)	0.00% (0/11)
MARITAL_STATUS	63.64% (7/11)	0.00% (0/11)	36.36% (4/11)
OCCUPATION	86.67% (52/60)	8.33% (5/60)	5.00% (3/60)
VEHICLE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
HEALTH_INFORMATION	81.63% (40/49)	10.20% (5/49)	8.16% (4/49)
EDUCATIONAL_RECORD	90.00% (9/10)	10.00% (1/10)	0.00% (0/10)
AGE	78.57% (44/56)	14.29% (8/56)	7.14% (4/56)
GENDER	92.31% (12/13)	0.00% (0/13)	7.69% (1/13)
ETHNICITY	100.00% (5/5)	0.00% (0/5)	0.00% (0/5)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
IP_ADDRESS	66.67% (2/3)	0.00% (0/3)	33.33% (1/3)
RACE	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
Overall	80.17% (954/1190)	11.93% (142/1190)	7.90% (94/1190)

Table 40: Weighted Results per Type and Overall (Oracle for **ShareGPT90K**), Model: **gpt-4.1**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	93.60% (161/172)	5.81% (10/172)	0.58% (1/172)
AFFILIATION	95.91% (164/171)	2.92% (5/171)	1.17% (2/171)
TIME	77.65% (205/264)	10.61% (28/264)	11.74% (31/264)
URL	100.00% (20/20)	0.00% (0/20)	0.00% (0/20)
EMAIL	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
GEOLOCATION	68.81% (225/327)	18.04% (59/327)	13.15% (43/327)
RELIGION	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
FINANCIAL_INFORMATION	81.82% (9/11)	18.18% (2/11)	0.00% (0/11)
MARITAL_STATUS	90.91% (10/11)	0.00% (0/11)	9.09% (1/11)
OCCUPATION	88.33% (53/60)	8.33% (5/60)	3.33% (2/60)
VEHICLE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	87.50% (7/8)	12.50% (1/8)	0.00% (0/8)
HEALTH_INFORMATION	87.76% (43/49)	6.12% (3/49)	6.12% (3/49)
EDUCATIONAL_RECORD	100.00% (10/10)	0.00% (0/10)	0.00% (0/10)
AGE	91.07% (51/56)	5.36% (3/56)	3.57% (2/56)
GENDER	92.31% (12/13)	7.69% (1/13)	0.00% (0/13)
ETHNICITY	100.00% (5/5)	0.00% (0/5)	0.00% (0/5)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
IP_ADDRESS	100.00% (3/3)	0.00% (0/3)	0.00% (0/3)
RACE	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
Overall	82.94% (987/1190)	9.92% (118/1190)	7.14% (85/1190)

Table 41: Weighted Results per Type and Overall (Oracle for **ShareGPT90K**), Model: **gpt-5**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	80.81% (139/172)	9.88% (17/172)	9.30% (16/172)
AFFILIATION	94.15% (161/171)	2.92% (5/171)	2.92% (5/171)
TIME	76.14% (201/264)	8.33% (22/264)	15.53% (41/264)
URL	85.00% (17/20)	10.00% (2/20)	5.00% (1/20)
EMAIL	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
GEOLOCATION	63.00% (206/327)	18.35% (60/327)	18.65% (61/327)
RELIGION	0.00% (0/2)	0.00% (0/2)	100.00% (2/2)
FINANCIAL_INFORMATION	63.64% (7/11)	27.27% (3/11)	9.09% (1/11)
MARITAL_STATUS	81.82% (9/11)	18.18% (2/11)	0.00% (0/11)
OCCUPATION	88.33% (53/60)	11.67% (7/60)	0.00% (0/60)
VEHICLE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	87.50% (7/8)	12.50% (1/8)	0.00% (0/8)
HEALTH_INFORMATION	65.31% (32/49)	12.24% (6/49)	22.45% (11/49)
EDUCATIONAL_RECORD	90.00% (9/10)	0.00% (0/10)	10.00% (1/10)
AGE	67.86% (38/56)	21.43% (12/56)	10.71% (6/56)
GENDER	76.92% (10/13)	15.38% (2/13)	7.69% (1/13)
ETHNICITY	60.00% (3/5)	0.00% (0/5)	40.00% (2/5)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
IP_ADDRESS	33.33% (1/3)	0.00% (0/3)	66.67% (2/3)
RACE	50.00% (1/2)	50.00% (1/2)	0.00% (0/2)
Overall	75.55% (899/1190)	11.85% (141/1190)	12.61% (150/1190)

Table 42: Weighted Results per Type and Overall (Oracle for **ShareGPT90K**), Model: **claude-3-7-sonnet-20250219**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	82.56% (142/172)	6.40% (11/172)	11.05% (19/172)
AFFILIATION	91.23% (156/171)	7.60% (13/171)	1.17% (2/171)
TIME	61.74% (163/264)	14.39% (38/264)	23.86% (63/264)
URL	75.00% (15/20)	15.00% (3/20)	10.00% (2/20)
EMAIL	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
GEOLOCATION	58.41% (191/327)	17.13% (56/327)	24.46% (80/327)
RELIGION	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
FINANCIAL_INFORMATION	63.64% (7/11)	36.36% (4/11)	0.00% (0/11)
MARITAL_STATUS	81.82% (9/11)	9.09% (1/11)	9.09% (1/11)
OCCUPATION	80.00% (48/60)	16.67% (10/60)	3.33% (2/60)
VEHICLE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	75.00% (6/8)	12.50% (1/8)	12.50% (1/8)
HEALTH_INFORMATION	65.31% (32/49)	12.24% (6/49)	22.45% (11/49)
EDUCATIONAL_RECORD	90.00% (9/10)	0.00% (0/10)	10.00% (1/10)
AGE	66.07% (37/56)	12.50% (7/56)	21.43% (12/56)
GENDER	69.23% (9/13)	7.69% (1/13)	23.08% (3/13)
ETHNICITY	60.00% (3/5)	40.00% (2/5)	0.00% (0/5)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
IP_ADDRESS	0.00% (0/3)	0.00% (0/3)	100.00% (3/3)
RACE	50.00% (1/2)	0.00% (0/2)	50.00% (1/2)
Overall	70.25% (836/1190)	12.86% (153/1190)	16.89% (201/1190)

Table 43: Weighted Results per Type and Overall (Oracle for **ShareGPT90K**), Model: **claude-sonnet-4-20250514**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	56.40% (97/172)	18.02% (31/172)	25.58% (44/172)
AFFILIATION	72.51% (124/171)	14.62% (25/171)	12.87% (22/171)
TIME	46.21% (122/264)	28.79% (76/264)	25.00% (66/264)
URL	60.00% (12/20)	25.00% (5/20)	15.00% (3/20)
EMAIL	75.00% (3/4)	0.00% (0/4)	25.00% (1/4)
GEOLOCATION	45.57% (149/327)	18.04% (59/327)	36.39% (119/327)
RELIGION	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
FINANCIAL_INFORMATION	36.36% (4/11)	27.27% (3/11)	36.36% (4/11)
MARITAL_STATUS	36.36% (4/11)	18.18% (2/11)	45.45% (5/11)
OCCUPATION	70.00% (42/60)	6.67% (4/60)	23.33% (14/60)
VEHICLE	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
INCOME	87.50% (7/8)	12.50% (1/8)	0.00% (0/8)
HEALTH_INFORMATION	63.27% (31/49)	10.20% (5/49)	26.53% (13/49)
EDUCATIONAL_RECORD	80.00% (8/10)	0.00% (0/10)	20.00% (2/10)
AGE	48.21% (27/56)	30.36% (17/56)	21.43% (12/56)
GENDER	76.92% (10/13)	0.00% (0/13)	23.08% (3/13)
ETHNICITY	60.00% (3/5)	20.00% (1/5)	20.00% (1/5)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
IP_ADDRESS	33.33% (1/3)	0.00% (0/3)	66.67% (2/3)
RACE	50.00% (1/2)	0.00% (0/2)	50.00% (1/2)
Overall	54.29% (646/1190)	19.41% (231/1190)	26.30% (313/1190)

Table 44: Weighted Results per Type and Overall (Oracle for **ShareGPT90K**), Model: **lgai/exaone-deep-32b**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	79.07% (136/172)	6.98% (12/172)	13.95% (24/172)
AFFILIATION	89.47% (153/171)	7.02% (12/171)	3.51% (6/171)
TIME	60.98% (161/264)	19.70% (52/264)	19.32% (51/264)
URL	95.00% (19/20)	0.00% (0/20)	5.00% (1/20)
EMAIL	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
GEOLOCATION	55.05% (180/327)	22.32% (73/327)	22.63% (74/327)
RELIGION	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
FINANCIAL_INFORMATION	54.55% (6/11)	18.18% (2/11)	27.27% (3/11)
MARITAL_STATUS	63.64% (7/11)	18.18% (2/11)	18.18% (2/11)
OCCUPATION	86.67% (52/60)	10.00% (6/60)	3.33% (2/60)
VEHICLE	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
INCOME	87.50% (7/8)	12.50% (1/8)	0.00% (0/8)
HEALTH_INFORMATION	63.27% (31/49)	22.45% (11/49)	14.29% (7/49)
EDUCATIONAL_RECORD	90.00% (9/10)	10.00% (1/10)	0.00% (0/10)
AGE	62.50% (35/56)	19.64% (11/56)	17.86% (10/56)
GENDER	61.54% (8/13)	15.38% (2/13)	23.08% (3/13)
ETHNICITY	60.00% (3/5)	40.00% (2/5)	0.00% (0/5)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
IP_ADDRESS	0.00% (0/3)	66.67% (2/3)	33.33% (1/3)
RACE	0.00% (0/2)	50.00% (1/2)	50.00% (1/2)
Overall	68.40% (814/1190)	15.97% (190/1190)	15.63% (186/1190)

Table 45: Weighted Results per Type and Overall (Oracle for **ShareGPT90K**), Model: **mistralai/mistral-small-3.1-24b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	75.00% (129/172)	9.88% (17/172)	15.12% (26/172)
AFFILIATION	81.87% (140/171)	11.70% (20/171)	6.43% (11/171)
TIME	52.65% (139/264)	15.53% (41/264)	31.82% (84/264)
URL	80.00% (16/20)	15.00% (3/20)	5.00% (1/20)
EMAIL	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
GEOLOCATION	53.21% (174/327)	18.96% (62/327)	27.83% (91/327)
RELIGION	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
FINANCIAL_INFORMATION	54.55% (6/11)	9.09% (1/11)	36.36% (4/11)
MARITAL_STATUS	54.55% (6/11)	9.09% (1/11)	36.36% (4/11)
OCCUPATION	65.00% (39/60)	15.00% (9/60)	20.00% (12/60)
VEHICLE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
HEALTH_INFORMATION	61.22% (30/49)	12.24% (6/49)	26.53% (13/49)
EDUCATIONAL_RECORD	90.00% (9/10)	0.00% (0/10)	10.00% (1/10)
AGE	53.57% (30/56)	23.21% (13/56)	23.21% (13/56)
GENDER	69.23% (9/13)	7.69% (1/13)	23.08% (3/13)
ETHNICITY	60.00% (3/5)	20.00% (1/5)	20.00% (1/5)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
IP_ADDRESS	33.33% (1/3)	33.33% (1/3)	33.33% (1/3)
RACE	50.00% (1/2)	0.00% (0/2)	50.00% (1/2)
Overall	62.86% (748/1190)	14.79% (176/1190)	22.35% (266/1190)

Table 46: Weighted Results per Type and Overall (Oracle for **ShareGPT90K**), Model: **qwen/qwen2.5-7b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	17.44% (30/172)	8.14% (14/172)	74.42% (128/172)
AFFILIATION	18.71% (32/171)	14.04% (24/171)	67.25% (115/171)
TIME	10.23% (27/264)	6.44% (17/264)	83.33% (220/264)
URL	30.00% (6/20)	0.00% (0/20)	70.00% (14/20)
EMAIL	0.00% (0/4)	25.00% (1/4)	75.00% (3/4)
GEOLOCATION	8.26% (27/327)	6.42% (21/327)	85.32% (279/327)
RELIGION	0.00% (0/2)	0.00% (0/2)	100.00% (2/2)
FINANCIAL_INFORMATION	9.09% (1/11)	0.00% (0/11)	90.91% (10/11)
MARITAL_STATUS	9.09% (1/11)	9.09% (1/11)	81.82% (9/11)
OCCUPATION	6.67% (4/60)	3.33% (2/60)	90.00% (54/60)
VEHICLE	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
INCOME	0.00% (0/8)	0.00% (0/8)	100.00% (8/8)
HEALTH_INFORMATION	22.45% (11/49)	12.24% (6/49)	65.31% (32/49)
EDUCATIONAL_RECORD	50.00% (5/10)	10.00% (1/10)	40.00% (4/10)
AGE	10.71% (6/56)	8.93% (5/56)	80.36% (45/56)
GENDER	7.69% (1/13)	7.69% (1/13)	84.62% (11/13)
ETHNICITY	0.00% (0/5)	20.00% (1/5)	80.00% (4/5)
ADDRESS	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
IP_ADDRESS	0.00% (0/3)	0.00% (0/3)	100.00% (3/3)
RACE	50.00% (1/2)	0.00% (0/2)	50.00% (1/2)
Overall	12.77% (152/1190)	7.90% (94/1190)	79.33% (944/1190)

Table 47: Weighted Results per Type and Overall (Oracle for **ShareGPT90K**), Model: **qwen/qwen2.5-0.5b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	69.77% (120/172)	29.07% (50/172)	1.16% (2/172)
AFFILIATION	53.80% (92/171)	45.03% (77/171)	1.17% (2/171)
TIME	45.83% (121/264)	53.41% (141/264)	0.76% (2/264)
URL	75.00% (15/20)	25.00% (5/20)	0.00% (0/20)
EMAIL	100.00% (4/4)	0.00% (0/4)	0.00% (0/4)
GEOLOCATION	42.51% (139/327)	57.49% (188/327)	0.00% (0/327)
RELIGION	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
FINANCIAL_INFORMATION	81.82% (9/11)	18.18% (2/11)	0.00% (0/11)
MARITAL_STATUS	63.64% (7/11)	36.36% (4/11)	0.00% (0/11)
OCCUPATION	71.67% (43/60)	21.67% (13/60)	6.67% (4/60)
VEHICLE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	62.50% (5/8)	37.50% (3/8)	0.00% (0/8)
HEALTH_INFORMATION	34.69% (17/49)	65.31% (32/49)	0.00% (0/49)
EDUCATIONAL_RECORD	20.00% (2/10)	80.00% (8/10)	0.00% (0/10)
AGE	53.57% (30/56)	46.43% (26/56)	0.00% (0/56)
GENDER	38.46% (5/13)	61.54% (8/13)	0.00% (0/13)
ETHNICITY	80.00% (4/5)	20.00% (1/5)	0.00% (0/5)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
IP_ADDRESS	0.00% (0/3)	100.00% (3/3)	0.00% (0/3)
RACE	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
Overall	51.68% (615/1190)	47.48% (565/1190)	0.84% (10/1190)

Table 48: Weighted Results per Type and Overall (Prediction for **ShareGPT90K**), Model: **gpt-4.1-nano**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	1.74% (3/172)	97.67% (168/172)	0.58% (1/172)
AFFILIATION	0.00% (0/171)	98.83% (169/171)	1.17% (2/171)
TIME	0.00% (0/264)	100.00% (264/264)	0.00% (0/264)
URL	0.00% (0/20)	100.00% (20/20)	0.00% (0/20)
EMAIL	0.00% (0/4)	100.00% (4/4)	0.00% (0/4)
GEOLOCATION	0.00% (0/327)	91.13% (298/327)	8.87% (29/327)
RELIGION	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
FINANCIAL_INFORMATION	0.00% (0/11)	100.00% (11/11)	0.00% (0/11)
MARITAL_STATUS	0.00% (0/11)	100.00% (11/11)	0.00% (0/11)
OCCUPATION	0.00% (0/60)	96.67% (58/60)	3.33% (2/60)
VEHICLE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/8)	100.00% (8/8)	0.00% (0/8)
HEALTH_INFORMATION	0.00% (0/49)	79.59% (39/49)	20.41% (10/49)
EDUCATIONAL_RECORD	0.00% (0/10)	100.00% (10/10)	0.00% (0/10)
AGE	0.00% (0/56)	100.00% (56/56)	0.00% (0/56)
GENDER	0.00% (0/13)	100.00% (13/13)	0.00% (0/13)
ETHNICITY	0.00% (0/5)	100.00% (5/5)	0.00% (0/5)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
IP_ADDRESS	0.00% (0/3)	100.00% (3/3)	0.00% (0/3)
RACE	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
Overall	0.25% (3/1190)	96.05% (1143/1190)	3.70% (44/1190)

Table 49: Weighted Results per Type and Overall (Prediction for **ShareGPT90K**), Model: **gpt-4.1**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	28.49% (49/172)	62.21% (107/172)	9.30% (16/172)
AFFILIATION	11.70% (20/171)	64.91% (111/171)	23.39% (40/171)
TIME	8.71% (23/264)	55.68% (147/264)	35.61% (94/264)
URL	30.00% (6/20)	30.00% (6/20)	40.00% (8/20)
EMAIL	75.00% (3/4)	0.00% (0/4)	25.00% (1/4)
GEOLOCATION	6.73% (22/327)	41.59% (136/327)	51.68% (169/327)
RELIGION	50.00% (1/2)	0.00% (0/2)	50.00% (1/2)
FINANCIAL_INFORMATION	0.00% (0/11)	54.55% (6/11)	45.45% (5/11)
MARITAL_STATUS	18.18% (2/11)	81.82% (9/11)	0.00% (0/11)
OCCUPATION	1.67% (1/60)	75.00% (45/60)	23.33% (14/60)
VEHICLE	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
INCOME	0.00% (0/8)	62.50% (5/8)	37.50% (3/8)
HEALTH_INFORMATION	28.57% (14/49)	63.27% (31/49)	8.16% (4/49)
EDUCATIONAL_RECORD	0.00% (0/10)	100.00% (10/10)	0.00% (0/10)
AGE	23.21% (13/56)	67.86% (38/56)	8.93% (5/56)
GENDER	46.15% (6/13)	46.15% (6/13)	7.69% (1/13)
ETHNICITY	40.00% (2/5)	40.00% (2/5)	20.00% (1/5)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
IP_ADDRESS	0.00% (0/3)	100.00% (3/3)	0.00% (0/3)
RACE	50.00% (1/2)	50.00% (1/2)	0.00% (0/2)
Overall	13.78% (164/1190)	55.71% (663/1190)	30.50% (363/1190)

Table 50: Weighted Results per Type and Overall (Prediction for **ShareGPT90K**), Model: **gpt-5**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	2.91% (5/172)	76.74% (132/172)	20.35% (35/172)
AFFILIATION	3.51% (6/171)	34.50% (59/171)	61.99% (106/171)
TIME	1.14% (3/264)	20.45% (54/264)	78.41% (207/264)
URL	40.00% (8/20)	25.00% (5/20)	35.00% (7/20)
EMAIL	0.00% (0/4)	75.00% (3/4)	25.00% (1/4)
GEOLOCATION	0.92% (3/327)	22.63% (74/327)	76.45% (250/327)
RELIGION	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
FINANCIAL_INFORMATION	0.00% (0/11)	54.55% (6/11)	45.45% (5/11)
MARITAL_STATUS	0.00% (0/11)	81.82% (9/11)	18.18% (2/11)
OCCUPATION	0.00% (0/60)	11.67% (7/60)	88.33% (53/60)
VEHICLE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/8)	75.00% (6/8)	25.00% (2/8)
HEALTH_INFORMATION	0.00% (0/49)	20.41% (10/49)	79.59% (39/49)
EDUCATIONAL_RECORD	0.00% (0/10)	30.00% (3/10)	70.00% (7/10)
AGE	5.36% (3/56)	51.79% (29/56)	42.86% (24/56)
GENDER	0.00% (0/13)	38.46% (5/13)	61.54% (8/13)
ETHNICITY	20.00% (1/5)	60.00% (3/5)	20.00% (1/5)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
IP_ADDRESS	0.00% (0/3)	0.00% (0/3)	100.00% (3/3)
RACE	0.00% (0/2)	50.00% (1/2)	50.00% (1/2)
Overall	2.44% (29/1190)	34.45% (410/1190)	63.11% (751/1190)

Table 51: Weighted Results per Type and Overall (Prediction for **ShareGPT90K**), Model: **claude-3-7-sonnet-20250219**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	2.33% (4/172)	85.47% (147/172)	12.21% (21/172)
AFFILIATION	0.00% (0/171)	57.89% (99/171)	42.11% (72/171)
TIME	0.38% (1/264)	48.48% (128/264)	51.14% (135/264)
URL	0.00% (0/20)	70.00% (14/20)	30.00% (6/20)
EMAIL	0.00% (0/4)	75.00% (3/4)	25.00% (1/4)
GEOLOCATION	0.00% (0/327)	35.78% (117/327)	64.22% (210/327)
RELIGION	0.00% (0/2)	0.00% (0/2)	100.00% (2/2)
FINANCIAL_INFORMATION	0.00% (0/11)	45.45% (5/11)	54.55% (6/11)
MARITAL_STATUS	0.00% (0/11)	100.00% (11/11)	0.00% (0/11)
OCCUPATION	0.00% (0/60)	45.00% (27/60)	55.00% (33/60)
VEHICLE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/8)	100.00% (8/8)	0.00% (0/8)
HEALTH_INFORMATION	0.00% (0/49)	20.41% (10/49)	79.59% (39/49)
EDUCATIONAL_RECORD	0.00% (0/10)	100.00% (10/10)	0.00% (0/10)
AGE	10.71% (6/56)	78.57% (44/56)	10.71% (6/56)
GENDER	7.69% (1/13)	92.31% (12/13)	0.00% (0/13)
ETHNICITY	20.00% (1/5)	80.00% (4/5)	0.00% (0/5)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
IP_ADDRESS	0.00% (0/3)	100.00% (3/3)	0.00% (0/3)
RACE	50.00% (1/2)	50.00% (1/2)	0.00% (0/2)
Overall	1.18% (14/1190)	54.20% (645/1190)	44.62% (531/1190)

Table 52: Weighted Results per Type and Overall (Prediction for **ShareGPT90K**), Model: **claude-sonnet-4-20250514**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	34.88% (60/172)	65.12% (112/172)	0.00% (0/172)
AFFILIATION	26.32% (45/171)	72.51% (124/171)	1.17% (2/171)
TIME	14.77% (39/264)	81.82% (216/264)	3.41% (9/264)
URL	25.00% (5/20)	75.00% (15/20)	0.00% (0/20)
EMAIL	75.00% (3/4)	25.00% (1/4)	0.00% (0/4)
GEOLOCATION	15.60% (51/327)	81.65% (267/327)	2.75% (9/327)
RELIGION	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
FINANCIAL_INFORMATION	9.09% (1/11)	90.91% (10/11)	0.00% (0/11)
MARITAL_STATUS	0.00% (0/11)	100.00% (11/11)	0.00% (0/11)
OCCUPATION	28.33% (17/60)	71.67% (43/60)	0.00% (0/60)
VEHICLE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	25.00% (2/8)	75.00% (6/8)	0.00% (0/8)
HEALTH_INFORMATION	18.37% (9/49)	81.63% (40/49)	0.00% (0/49)
EDUCATIONAL_RECORD	20.00% (2/10)	80.00% (8/10)	0.00% (0/10)
AGE	23.21% (13/56)	67.86% (38/56)	8.93% (5/56)
GENDER	46.15% (6/13)	53.85% (7/13)	0.00% (0/13)
ETHNICITY	0.00% (0/5)	100.00% (5/5)	0.00% (0/5)
ADDRESS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
IP_ADDRESS	100.00% (3/3)	0.00% (0/3)	0.00% (0/3)
RACE	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
Overall	21.76% (259/1190)	76.13% (906/1190)	2.10% (25/1190)

Table 53: Weighted Results per Type and Overall (Prediction for **ShareGPT90K**), Model: **lgai/exaone-deep-32b**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	0.58% (1/172)	98.84% (170/172)	0.58% (1/172)
AFFILIATION	0.00% (0/171)	98.25% (168/171)	1.75% (3/171)
TIME	0.00% (0/264)	98.11% (259/264)	1.89% (5/264)
URL	15.00% (3/20)	75.00% (15/20)	10.00% (2/20)
EMAIL	0.00% (0/4)	100.00% (4/4)	0.00% (0/4)
GEOLOCATION	0.31% (1/327)	97.55% (319/327)	2.14% (7/327)
RELIGION	0.00% (0/2)	50.00% (1/2)	50.00% (1/2)
FINANCIAL_INFORMATION	9.09% (1/11)	81.82% (9/11)	9.09% (1/11)
MARITAL_STATUS	0.00% (0/11)	100.00% (11/11)	0.00% (0/11)
OCCUPATION	0.00% (0/60)	98.33% (59/60)	1.67% (1/60)
VEHICLE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/8)	100.00% (8/8)	0.00% (0/8)
HEALTH_INFORMATION	0.00% (0/49)	87.76% (43/49)	12.24% (6/49)
EDUCATIONAL_RECORD	0.00% (0/10)	80.00% (8/10)	20.00% (2/10)
AGE	0.00% (0/56)	100.00% (56/56)	0.00% (0/56)
GENDER	0.00% (0/13)	100.00% (13/13)	0.00% (0/13)
ETHNICITY	0.00% (0/5)	100.00% (5/5)	0.00% (0/5)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
IP_ADDRESS	0.00% (0/3)	100.00% (3/3)	0.00% (0/3)
RACE	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
Overall	0.50% (6/1190)	97.06% (1155/1190)	2.44% (29/1190)

Table 54: Weighted Results per Type and Overall (Prediction for **ShareGPT90K**), Model: **mistralai/mistral-small-3.1-24b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	0.58% (1/172)	99.42% (171/172)	0.00% (0/172)
AFFILIATION	1.17% (2/171)	98.83% (169/171)	0.00% (0/171)
TIME	1.14% (3/264)	98.86% (261/264)	0.00% (0/264)
URL	10.00% (2/20)	90.00% (18/20)	0.00% (0/20)
EMAIL	75.00% (3/4)	25.00% (1/4)	0.00% (0/4)
GEOLOCATION	0.61% (2/327)	99.39% (325/327)	0.00% (0/327)
RELIGION	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
FINANCIAL_INFORMATION	18.18% (2/11)	81.82% (9/11)	0.00% (0/11)
MARITAL_STATUS	9.09% (1/11)	90.91% (10/11)	0.00% (0/11)
OCCUPATION	0.00% (0/60)	100.00% (60/60)	0.00% (0/60)
VEHICLE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/8)	100.00% (8/8)	0.00% (0/8)
HEALTH_INFORMATION	0.00% (0/49)	100.00% (49/49)	0.00% (0/49)
EDUCATIONAL_RECORD	0.00% (0/10)	100.00% (10/10)	0.00% (0/10)
AGE	0.00% (0/56)	100.00% (56/56)	0.00% (0/56)
GENDER	0.00% (0/13)	100.00% (13/13)	0.00% (0/13)
ETHNICITY	0.00% (0/5)	100.00% (5/5)	0.00% (0/5)
ADDRESS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
IP_ADDRESS	0.00% (0/3)	100.00% (3/3)	0.00% (0/3)
RACE	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
Overall	1.34% (16/1190)	98.66% (1174/1190)	0.00% (0/1190)

Table 55: Weighted Results per Type and Overall (Prediction for **ShareGPT90K**), Model: **qwen/qwen2.5-7b-instruct**

H.4 WILDCHAT

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	88.82% (135/152)	3.95% (6/152)	7.24% (11/152)
AFFILIATION	90.91% (130/143)	6.99% (10/143)	2.10% (3/143)
GEOLOCATION	84.39% (200/237)	6.75% (16/237)	8.86% (21/237)
USERNAME	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
TIME	85.23% (127/149)	8.05% (12/149)	6.71% (10/149)
AGE	63.64% (14/22)	18.18% (4/22)	18.18% (4/22)
OCCUPATION	81.08% (30/37)	10.81% (4/37)	8.11% (3/37)
QUANTITY	100.00% (6/6)	0.00% (0/6)	0.00% (0/6)
ETHNICITY	77.78% (7/9)	11.11% (1/9)	11.11% (1/9)
GENDER	85.71% (6/7)	14.29% (1/7)	0.00% (0/7)
EMAIL	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
URL	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
HEALTH_INFORMATION	50.00% (3/6)	33.33% (2/6)	16.67% (1/6)
RACE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	85.71% (12/14)	14.29% (2/14)	0.00% (0/14)
PRODUCT	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	83.33% (5/6)	0.00% (0/6)	16.67% (1/6)
PHONE_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
EDUCATIONAL_RECORD	100.00% (11/11)	0.00% (0/11)	0.00% (0/11)
ID_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	86.20% (706/819)	7.08% (58/819)	6.72% (55/819)

Table 56: Weighted Results per Type and Overall (Oracle for **WildChat**), Model: **gpt-4.1-nano**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	88.82% (135/152)	5.92% (9/152)	5.26% (8/152)
AFFILIATION	91.61% (131/143)	4.20% (6/143)	4.20% (6/143)
GEOLOCATION	81.01% (192/237)	9.28% (22/237)	9.70% (23/237)
USERNAME	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
TIME	85.23% (127/149)	6.71% (10/149)	8.05% (12/149)
AGE	72.73% (16/22)	9.09% (2/22)	18.18% (4/22)
OCCUPATION	89.19% (33/37)	5.41% (2/37)	5.41% (2/37)
QUANTITY	100.00% (6/6)	0.00% (0/6)	0.00% (0/6)
ETHNICITY	77.78% (7/9)	0.00% (0/9)	22.22% (2/9)
GENDER	71.43% (5/7)	14.29% (1/7)	14.29% (1/7)
EMAIL	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
URL	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
HEALTH_INFORMATION	83.33% (5/6)	16.67% (1/6)	0.00% (0/6)
RACE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (14/14)	0.00% (0/14)	0.00% (0/14)
PRODUCT	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	83.33% (5/6)	16.67% (1/6)	0.00% (0/6)
PHONE_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
EDUCATIONAL_RECORD	81.82% (9/11)	18.18% (2/11)	0.00% (0/11)
ID_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	86.08% (705/819)	6.84% (56/819)	7.08% (58/819)

Table 57: Weighted Results per Type and Overall (Oracle for **WildChat**), Model: **gpt-4.1**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	89.47% (136/152)	7.89% (12/152)	2.63% (4/152)
AFFILIATION	93.71% (134/143)	4.90% (7/143)	1.40% (2/143)
GEOLOCATION	86.50% (205/237)	9.70% (23/237)	3.80% (9/237)
USERNAME	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
TIME	91.28% (136/149)	4.03% (6/149)	4.70% (7/149)
AGE	77.27% (17/22)	13.64% (3/22)	9.09% (2/22)
OCCUPATION	86.49% (32/37)	8.11% (3/37)	5.41% (2/37)
QUANTITY	100.00% (6/6)	0.00% (0/6)	0.00% (0/6)
ETHNICITY	77.78% (7/9)	11.11% (1/9)	11.11% (1/9)
GENDER	85.71% (6/7)	0.00% (0/7)	14.29% (1/7)
EMAIL	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
URL	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
HEALTH_INFORMATION	100.00% (6/6)	0.00% (0/6)	0.00% (0/6)
RACE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (14/14)	0.00% (0/14)	0.00% (0/14)
PRODUCT	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	83.33% (5/6)	0.00% (0/6)	16.67% (1/6)
PHONE_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
EDUCATIONAL_RECORD	100.00% (11/11)	0.00% (0/11)	0.00% (0/11)
ID_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	89.74% (735/819)	6.72% (55/819)	3.54% (29/819)

Table 58: Weighted Results per Type and Overall (Oracle for **WildChat**), Model: **gpt-5**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	77.63% (118/152)	9.87% (15/152)	12.50% (19/152)
AFFILIATION	83.92% (120/143)	7.69% (11/143)	8.39% (12/143)
GEOLOCATION	78.06% (185/237)	9.28% (22/237)	12.66% (30/237)
USERNAME	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
TIME	83.89% (125/149)	10.74% (16/149)	5.37% (8/149)
AGE	63.64% (14/22)	13.64% (3/22)	22.73% (5/22)
OCCUPATION	78.38% (29/37)	5.41% (2/37)	16.22% (6/37)
QUANTITY	100.00% (6/6)	0.00% (0/6)	0.00% (0/6)
ETHNICITY	33.33% (3/9)	0.00% (0/9)	66.67% (6/9)
GENDER	85.71% (6/7)	0.00% (0/7)	14.29% (1/7)
EMAIL	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
URL	75.00% (6/8)	12.50% (1/8)	12.50% (1/8)
HEALTH_INFORMATION	100.00% (6/6)	0.00% (0/6)	0.00% (0/6)
RACE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	100.00% (14/14)	0.00% (0/14)	0.00% (0/14)
PRODUCT	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	66.67% (4/6)	16.67% (1/6)	16.67% (1/6)
PHONE_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
EDUCATIONAL_RECORD	100.00% (11/11)	0.00% (0/11)	0.00% (0/11)
ID_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	80.34% (658/819)	8.79% (72/819)	10.87% (89/819)

Table 59: Weighted Results per Type and Overall (Oracle for **WildChat**), Model: **claude-3-7-sonnet-20250219**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	81.58% (124/152)	7.24% (11/152)	11.18% (17/152)
AFFILIATION	85.31% (122/143)	6.29% (9/143)	8.39% (12/143)
GEOLOCATION	78.90% (187/237)	9.70% (23/237)	11.39% (27/237)
USERNAME	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
TIME	81.21% (121/149)	12.08% (18/149)	6.71% (10/149)
AGE	63.64% (14/22)	13.64% (3/22)	22.73% (5/22)
OCCUPATION	83.78% (31/37)	5.41% (2/37)	10.81% (4/37)
QUANTITY	100.00% (6/6)	0.00% (0/6)	0.00% (0/6)
ETHNICITY	44.44% (4/9)	0.00% (0/9)	55.56% (5/9)
GENDER	71.43% (5/7)	28.57% (2/7)	0.00% (0/7)
EMAIL	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
URL	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
HEALTH_INFORMATION	83.33% (5/6)	16.67% (1/6)	0.00% (0/6)
RACE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	92.86% (13/14)	0.00% (0/14)	7.14% (1/14)
PRODUCT	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	83.33% (5/6)	16.67% (1/6)	0.00% (0/6)
PHONE_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
EDUCATIONAL_RECORD	81.82% (9/11)	18.18% (2/11)	0.00% (0/11)
ID_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	81.32% (666/819)	8.79% (72/819)	9.89% (81/819)

Table 60: Weighted Results per Type and Overall (Oracle for **WildChat**), Model: **claude-sonnet-4-20250514**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	67.76% (103/152)	18.42% (28/152)	13.82% (21/152)
AFFILIATION	73.43% (105/143)	12.59% (18/143)	13.99% (20/143)
GEOLOCATION	67.09% (159/237)	12.24% (29/237)	20.68% (49/237)
USERNAME	50.00% (1/2)	50.00% (1/2)	0.00% (0/2)
TIME	65.10% (97/149)	18.12% (27/149)	16.78% (25/149)
AGE	63.64% (14/22)	13.64% (3/22)	22.73% (5/22)
OCCUPATION	72.97% (27/37)	10.81% (4/37)	16.22% (6/37)
QUANTITY	66.67% (4/6)	16.67% (1/6)	16.67% (1/6)
ETHNICITY	55.56% (5/9)	22.22% (2/9)	22.22% (2/9)
GENDER	71.43% (5/7)	14.29% (1/7)	14.29% (1/7)
EMAIL	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
URL	87.50% (7/8)	12.50% (1/8)	0.00% (0/8)
HEALTH_INFORMATION	66.67% (4/6)	16.67% (1/6)	16.67% (1/6)
RACE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	92.86% (13/14)	0.00% (0/14)	7.14% (1/14)
PRODUCT	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	66.67% (4/6)	16.67% (1/6)	16.67% (1/6)
PHONE_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
EDUCATIONAL_RECORD	81.82% (9/11)	9.09% (1/11)	9.09% (1/11)
ID_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	69.23% (567/819)	14.41% (118/819)	16.36% (134/819)

Table 61: Weighted Results per Type and Overall (Oracle for **WildChat**), Model: **lgai/exaone-deep-32b**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	83.55% (127/152)	8.55% (13/152)	7.89% (12/152)
AFFILIATION	90.91% (130/143)	5.59% (8/143)	3.50% (5/143)
GEOLOCATION	83.54% (198/237)	7.17% (17/237)	9.28% (22/237)
USERNAME	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
TIME	83.22% (124/149)	10.07% (15/149)	6.71% (10/149)
AGE	77.27% (17/22)	9.09% (2/22)	13.64% (3/22)
OCCUPATION	86.49% (32/37)	2.70% (1/37)	10.81% (4/37)
QUANTITY	100.00% (6/6)	0.00% (0/6)	0.00% (0/6)
ETHNICITY	66.67% (6/9)	11.11% (1/9)	22.22% (2/9)
GENDER	71.43% (5/7)	14.29% (1/7)	14.29% (1/7)
EMAIL	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
URL	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
HEALTH_INFORMATION	66.67% (4/6)	33.33% (2/6)	0.00% (0/6)
RACE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	100.00% (14/14)	0.00% (0/14)	0.00% (0/14)
PRODUCT	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	66.67% (4/6)	16.67% (1/6)	16.67% (1/6)
PHONE_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
EDUCATIONAL_RECORD	100.00% (11/11)	0.00% (0/11)	0.00% (0/11)
ID_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	85.23% (698/819)	7.45% (61/819)	7.33% (60/819)

Table 62: Weighted Results per Type and Overall (Oracle for **WildChat**), Model: **mistralai/mistral-small-3.1-24b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	83.55% (127/152)	5.26% (8/152)	11.18% (17/152)
AFFILIATION	88.11% (126/143)	4.90% (7/143)	6.99% (10/143)
GEOLOCATION	78.06% (185/237)	9.70% (23/237)	12.24% (29/237)
USERNAME	50.00% (1/2)	0.00% (0/2)	50.00% (1/2)
TIME	75.17% (112/149)	10.07% (15/149)	14.77% (22/149)
AGE	81.82% (18/22)	0.00% (0/22)	18.18% (4/22)
OCCUPATION	59.46% (22/37)	13.51% (5/37)	27.03% (10/37)
QUANTITY	100.00% (6/6)	0.00% (0/6)	0.00% (0/6)
ETHNICITY	66.67% (6/9)	0.00% (0/9)	33.33% (3/9)
GENDER	71.43% (5/7)	14.29% (1/7)	14.29% (1/7)
EMAIL	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
URL	100.00% (8/8)	0.00% (0/8)	0.00% (0/8)
HEALTH_INFORMATION	83.33% (5/6)	16.67% (1/6)	0.00% (0/6)
RACE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	78.57% (11/14)	21.43% (3/14)	0.00% (0/14)
PRODUCT	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	83.33% (5/6)	0.00% (0/6)	16.67% (1/6)
PHONE_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
EDUCATIONAL_RECORD	81.82% (9/11)	18.18% (2/11)	0.00% (0/11)
ID_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
Overall	80.10% (656/819)	7.94% (65/819)	11.97% (98/819)

Table 63: Weighted Results per Type and Overall (Oracle for **WildChat**), Model: **qwen/qwen2.5-7b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	36.18% (55/152)	10.53% (16/152)	53.29% (81/152)
AFFILIATION	36.36% (52/143)	16.08% (23/143)	47.55% (68/143)
GEOLOCATION	24.05% (57/237)	21.10% (50/237)	54.85% (130/237)
USERNAME	0.00% (0/2)	0.00% (0/2)	100.00% (2/2)
TIME	25.50% (38/149)	13.42% (20/149)	61.07% (91/149)
AGE	9.09% (2/22)	9.09% (2/22)	81.82% (18/22)
OCCUPATION	29.73% (11/37)	8.11% (3/37)	62.16% (23/37)
QUANTITY	50.00% (3/6)	50.00% (3/6)	0.00% (0/6)
ETHNICITY	11.11% (1/9)	0.00% (0/9)	88.89% (8/9)
GENDER	42.86% (3/7)	28.57% (2/7)	28.57% (2/7)
EMAIL	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
URL	50.00% (4/8)	25.00% (2/8)	25.00% (2/8)
HEALTH_INFORMATION	16.67% (1/6)	0.00% (0/6)	83.33% (5/6)
RACE	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
INCOME	7.14% (1/14)	0.00% (0/14)	92.86% (13/14)
PRODUCT	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
FINANCIAL_INFORMATION	33.33% (2/6)	50.00% (3/6)	16.67% (1/6)
PHONE_NUMBER	50.00% (1/2)	0.00% (0/2)	50.00% (1/2)
EDUCATIONAL_RECORD	18.18% (2/11)	0.00% (0/11)	81.82% (9/11)
ID_NUMBER	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
Overall	28.82% (236/819)	15.38% (126/819)	55.80% (457/819)

Table 64: Weighted Results per Type and Overall (Oracle for **WildChat**), Model: **qwen/qwen2.5-0.5b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	64.47% (98/152)	34.21% (52/152)	1.32% (2/152)
AFFILIATION	44.76% (64/143)	55.24% (79/143)	0.00% (0/143)
GEOLOCATION	59.07% (140/237)	40.93% (97/237)	0.00% (0/237)
USERNAME	50.00% (1/2)	50.00% (1/2)	0.00% (0/2)
TIME	40.94% (61/149)	59.06% (88/149)	0.00% (0/149)
AGE	63.64% (14/22)	36.36% (8/22)	0.00% (0/22)
OCCUPATION	45.95% (17/37)	54.05% (20/37)	0.00% (0/37)
QUANTITY	50.00% (3/6)	50.00% (3/6)	0.00% (0/6)
ETHNICITY	55.56% (5/9)	44.44% (4/9)	0.00% (0/9)
GENDER	71.43% (5/7)	28.57% (2/7)	0.00% (0/7)
EMAIL	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
URL	87.50% (7/8)	12.50% (1/8)	0.00% (0/8)
HEALTH_INFORMATION	83.33% (5/6)	16.67% (1/6)	0.00% (0/6)
RACE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	92.86% (13/14)	7.14% (1/14)	0.00% (0/14)
PRODUCT	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
FINANCIAL_INFORMATION	66.67% (4/6)	33.33% (2/6)	0.00% (0/6)
PHONE_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
EDUCATIONAL_RECORD	0.00% (0/11)	100.00% (11/11)	0.00% (0/11)
ID_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	54.46% (446/819)	45.30% (371/819)	0.24% (2/819)

Table 65: Weighted Results per Type and Overall (Prediction for **WildChat**), Model: **gpt-4.1-nano**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	0.00% (0/152)	100.00% (152/152)	0.00% (0/152)
AFFILIATION	0.00% (0/143)	100.00% (143/143)	0.00% (0/143)
GEOLOCATION	0.00% (0/237)	92.83% (220/237)	7.17% (17/237)
USERNAME	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
TIME	0.00% (0/149)	98.66% (147/149)	1.34% (2/149)
AGE	0.00% (0/22)	95.45% (21/22)	4.55% (1/22)
OCCUPATION	0.00% (0/37)	100.00% (37/37)	0.00% (0/37)
QUANTITY	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
ETHNICITY	0.00% (0/9)	100.00% (9/9)	0.00% (0/9)
GENDER	0.00% (0/7)	100.00% (7/7)	0.00% (0/7)
EMAIL	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
URL	0.00% (0/8)	100.00% (8/8)	0.00% (0/8)
HEALTH_INFORMATION	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
RACE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/14)	100.00% (14/14)	0.00% (0/14)
PRODUCT	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
PHONE_NUMBER	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
EDUCATIONAL_RECORD	0.00% (0/11)	100.00% (11/11)	0.00% (0/11)
ID_NUMBER	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
KEYS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
GPA	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	0.00% (0/819)	97.56% (799/819)	2.44% (20/819)

Table 66: Weighted Results per Type and Overall (Prediction for **WildChat**), Model: **gpt-4.1**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	16.45% (25/152)	57.89% (88/152)	25.66% (39/152)
AFFILIATION	11.19% (16/143)	59.44% (85/143)	29.37% (42/143)
GEOLOCATION	8.44% (20/237)	57.38% (136/237)	34.18% (81/237)
USERNAME	50.00% (1/2)	50.00% (1/2)	0.00% (0/2)
TIME	19.46% (29/149)	68.46% (102/149)	12.08% (18/149)
AGE	13.64% (3/22)	72.73% (16/22)	13.64% (3/22)
OCCUPATION	2.70% (1/37)	75.68% (28/37)	21.62% (8/37)
QUANTITY	0.00% (0/6)	16.67% (1/6)	83.33% (5/6)
ETHNICITY	0.00% (0/9)	55.56% (5/9)	44.44% (4/9)
GENDER	28.57% (2/7)	57.14% (4/7)	14.29% (1/7)
EMAIL	0.00% (0/2)	0.00% (0/2)	100.00% (2/2)
URL	12.50% (1/8)	50.00% (4/8)	37.50% (3/8)
HEALTH_INFORMATION	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
RACE	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
INCOME	14.29% (2/14)	85.71% (12/14)	0.00% (0/14)
PRODUCT	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
FINANCIAL_INFORMATION	0.00% (0/6)	16.67% (1/6)	83.33% (5/6)
PHONE_NUMBER	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
EDUCATIONAL_RECORD	0.00% (0/11)	81.82% (9/11)	18.18% (2/11)
ID_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
KEYS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
GPA	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	12.45% (102/819)	61.29% (502/819)	26.25% (215/819)

Table 67: Weighted Results per Type and Overall (Prediction for **WildChat**), Model: **gpt-5**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	1.97% (3/152)	61.84% (94/152)	36.18% (55/152)
AFFILIATION	0.00% (0/143)	38.46% (55/143)	61.54% (88/143)
GEOLOCATION	0.42% (1/237)	27.00% (64/237)	72.57% (172/237)
USERNAME	50.00% (1/2)	50.00% (1/2)	0.00% (0/2)
TIME	3.36% (5/149)	27.52% (41/149)	69.13% (103/149)
AGE	4.55% (1/22)	45.45% (10/22)	50.00% (11/22)
OCCUPATION	0.00% (0/37)	32.43% (12/37)	67.57% (25/37)
QUANTITY	0.00% (0/6)	0.00% (0/6)	100.00% (6/6)
ETHNICITY	0.00% (0/9)	11.11% (1/9)	88.89% (8/9)
GENDER	0.00% (0/7)	0.00% (0/7)	100.00% (7/7)
EMAIL	0.00% (0/2)	50.00% (1/2)	50.00% (1/2)
URL	12.50% (1/8)	37.50% (3/8)	50.00% (4/8)
HEALTH_INFORMATION	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
RACE	0.00% (0/1)	0.00% (0/1)	100.00% (1/1)
INCOME	0.00% (0/14)	28.57% (4/14)	71.43% (10/14)
PRODUCT	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	0.00% (0/6)	66.67% (4/6)	33.33% (2/6)
PHONE_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
EDUCATIONAL_RECORD	0.00% (0/11)	81.82% (9/11)	18.18% (2/11)
ID_NUMBER	0.00% (0/2)	0.00% (0/2)	100.00% (2/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	1.83% (15/819)	37.48% (307/819)	60.68% (497/819)

Table 68: Weighted Results per Type and Overall (Prediction for **WildChat**), Model: **claude-3-7-sonnet-20250219**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	0.66% (1/152)	76.97% (117/152)	22.37% (34/152)
AFFILIATION	0.00% (0/143)	60.84% (87/143)	39.16% (56/143)
GEOLOCATION	0.42% (1/237)	34.60% (82/237)	64.98% (154/237)
USERNAME	50.00% (1/2)	50.00% (1/2)	0.00% (0/2)
TIME	0.00% (0/149)	57.05% (85/149)	42.95% (64/149)
AGE	0.00% (0/22)	81.82% (18/22)	18.18% (4/22)
OCCUPATION	0.00% (0/37)	62.16% (23/37)	37.84% (14/37)
QUANTITY	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
ETHNICITY	0.00% (0/9)	66.67% (6/9)	33.33% (3/9)
GENDER	0.00% (0/7)	42.86% (3/7)	57.14% (4/7)
EMAIL	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
URL	12.50% (1/8)	75.00% (6/8)	12.50% (1/8)
HEALTH_INFORMATION	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
RACE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/14)	100.00% (14/14)	0.00% (0/14)
PRODUCT	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	0.00% (0/6)	66.67% (4/6)	33.33% (2/6)
PHONE_NUMBER	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
EDUCATIONAL_RECORD	0.00% (0/11)	100.00% (11/11)	0.00% (0/11)
ID_NUMBER	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
KEYS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
GPA	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	0.49% (4/819)	58.49% (479/819)	41.03% (336/819)

Table 69: Weighted Results per Type and Overall (Prediction for **WildChat**), Model: **claude-sonnet-4-20250514**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	51.97% (79/152)	47.37% (72/152)	0.66% (1/152)
AFFILIATION	34.27% (49/143)	65.73% (94/143)	0.00% (0/143)
GEOLOCATION	35.86% (85/237)	64.14% (152/237)	0.00% (0/237)
USERNAME	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
TIME	40.94% (61/149)	57.72% (86/149)	1.34% (2/149)
AGE	31.82% (7/22)	63.64% (14/22)	4.55% (1/22)
OCCUPATION	13.51% (5/37)	86.49% (32/37)	0.00% (0/37)
QUANTITY	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
ETHNICITY	66.67% (6/9)	33.33% (3/9)	0.00% (0/9)
GENDER	71.43% (5/7)	28.57% (2/7)	0.00% (0/7)
EMAIL	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
URL	75.00% (6/8)	25.00% (2/8)	0.00% (0/8)
HEALTH_INFORMATION	16.67% (1/6)	83.33% (5/6)	0.00% (0/6)
RACE	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
INCOME	0.00% (0/14)	100.00% (14/14)	0.00% (0/14)
PRODUCT	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	33.33% (2/6)	66.67% (4/6)	0.00% (0/6)
PHONE_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
EDUCATIONAL_RECORD	18.18% (2/11)	81.82% (9/11)	0.00% (0/11)
ID_NUMBER	100.00% (2/2)	0.00% (0/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	38.58% (316/819)	60.93% (499/819)	0.49% (4/819)

Table 70: Weighted Results per Type and Overall (Prediction for **WildChat**), Model: **lgai/exaone-deep-32b**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	1.97% (3/152)	97.37% (148/152)	0.66% (1/152)
AFFILIATION	0.00% (0/143)	98.60% (141/143)	1.40% (2/143)
GEOLOCATION	0.00% (0/237)	99.58% (236/237)	0.42% (1/237)
USERNAME	50.00% (1/2)	50.00% (1/2)	0.00% (0/2)
TIME	0.00% (0/149)	97.99% (146/149)	2.01% (3/149)
AGE	4.55% (1/22)	95.45% (21/22)	0.00% (0/22)
OCCUPATION	0.00% (0/37)	94.59% (35/37)	5.41% (2/37)
QUANTITY	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
ETHNICITY	0.00% (0/9)	100.00% (9/9)	0.00% (0/9)
GENDER	0.00% (0/7)	100.00% (7/7)	0.00% (0/7)
EMAIL	0.00% (0/2)	50.00% (1/2)	50.00% (1/2)
URL	12.50% (1/8)	87.50% (7/8)	0.00% (0/8)
HEALTH_INFORMATION	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
RACE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/14)	100.00% (14/14)	0.00% (0/14)
PRODUCT	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
PHONE_NUMBER	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
EDUCATIONAL_RECORD	0.00% (0/11)	27.27% (3/11)	72.73% (8/11)
ID_NUMBER	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
KEYS	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
GPA	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	0.73% (6/819)	97.07% (795/819)	2.20% (18/819)

Table 71: Weighted Results per Type and Overall (Prediction for **WildChat**), Model: **mistralai/mistral-small-3.1-24b-instruct**

Type	Weighted Redact	Weighted Abstract	Weighted Retain
NAME	3.95% (6/152)	96.05% (146/152)	0.00% (0/152)
AFFILIATION	0.00% (0/143)	97.90% (140/143)	2.10% (3/143)
GEOLOCATION	1.27% (3/237)	98.73% (234/237)	0.00% (0/237)
USERNAME	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
TIME	4.03% (6/149)	95.30% (142/149)	0.67% (1/149)
AGE	0.00% (0/22)	100.00% (22/22)	0.00% (0/22)
OCCUPATION	0.00% (0/37)	100.00% (37/37)	0.00% (0/37)
QUANTITY	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
ETHNICITY	0.00% (0/9)	100.00% (9/9)	0.00% (0/9)
GENDER	0.00% (0/7)	100.00% (7/7)	0.00% (0/7)
EMAIL	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
URL	0.00% (0/8)	100.00% (8/8)	0.00% (0/8)
HEALTH_INFORMATION	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
RACE	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
INCOME	0.00% (0/14)	100.00% (14/14)	0.00% (0/14)
PRODUCT	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
FINANCIAL_INFORMATION	0.00% (0/6)	100.00% (6/6)	0.00% (0/6)
PHONE_NUMBER	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
EDUCATIONAL_RECORD	0.00% (0/11)	100.00% (11/11)	0.00% (0/11)
ID_NUMBER	0.00% (0/2)	100.00% (2/2)	0.00% (0/2)
KEYS	100.00% (1/1)	0.00% (0/1)	0.00% (0/1)
GPA	0.00% (0/1)	100.00% (1/1)	0.00% (0/1)
Overall	1.95% (16/819)	97.56% (799/819)	0.49% (4/819)

Table 72: Weighted Results per Type and Overall (Prediction for **WildChat**), Model: **qwen/qwen2.5-7b-instruct**

I CROSS-MODEL DATA MINIMIZATION OVERLAP

This section details the cross-model overlap experiment from §6.1, which quantifies how similar data minimization decisions are across response generation models, and whether the minimal sufficient masking chosen by each model's oracle shows stable structure.

For each model and dataset, we compare the oracle-derived action map to the GPT-5 oracle (as reference), computing Jaccard overlap separately for REDACT and ABSTRACT. The overlap is the number of PII spans where both oracles choose the same action divided by the union of spans where either does, yielding model- and dataset-level consistency measures.

Tables 73 and 74 report the resulting overlaps. As summarized in §6.1, redaction overlap is high across most models and datasets, typically at or above eighty percent, and in some datasets such as CaseHOLD it reaches above ninety percent for nearly all models. These results indicate that the majority of removed spans form a shared core of non-essential sensitive information that models broadly agree upon. In contrast, abstraction overlap is lower but abstraction itself accounts for a much smaller fraction of actions, which makes its variability less consequential in practice. Taken together, these observations suggest that the essential privacy-preserving behavior, namely which spans can be safely removed while maintaining utility, generalizes well across model families even though the exact minimally sufficient prompts remain model specific by definition.

Dataset	Model	Overlap
ShareGPT	gpt-4.1-nano	0.802493
ShareGPT	gpt-4.1	0.845057
ShareGPT	claude-3-7-sonnet-20250219	0.792776
ShareGPT	claude-sonnet-4-20250514	0.754572
ShareGPT	lgai/exaone-deep-32b	0.583899
ShareGPT	mistralai/mistral-small-3.1-24b-instruct	0.743466
ShareGPT	qwen/qwen2.5-7b-instruct	0.686103
ShareGPT	qwen/qwen2.5-0.5b-instruct	0.145875
WildChat	gpt-4.1-nano	0.849807
WildChat	gpt-4.1	0.860465
WildChat	claude-3-7-sonnet-20250219	0.797419
WildChat	claude-sonnet-4-20250514	0.800771
WildChat	lgai/exaone-deep-32b	0.701961
WildChat	mistralai/mistral-small-3.1-24b-instruct	0.853816
WildChat	qwen/qwen2.5-7b-instruct	0.794839
WildChat	qwen/qwen2.5-0.5b-instruct	0.298128
MedQA	gpt-4.1-nano	0.875905
MedQA	gpt-4.1	0.950464
MedQA	claude-3-7-sonnet-20250219	0.708768
MedQA	claude-sonnet-4-20250514	0.952183
MedQA	lgai/exaone-deep-32b	0.762055
MedQA	mistralai/mistral-small-3.1-24b-instruct	0.954451
MedQA	qwen/qwen2.5-7b-instruct	0.889583
MedQA	qwen/qwen2.5-0.5b-instruct	0.292683
CaseHOLD	gpt-4.1-nano	0.945274
CaseHOLD	gpt-4.1	0.985075
CaseHOLD	claude-3-7-sonnet-20250219	0.987531
CaseHOLD	claude-sonnet-4-20250514	0.987562
CaseHOLD	lgai/exaone-deep-32b	0.736181
CaseHOLD	mistralai/mistral-small-3.1-24b-instruct	0.937811
CaseHOLD	qwen/qwen2.5-7b-instruct	0.957711
CaseHOLD	qwen/qwen2.5-0.5b-instruct	0.395522

Table 73: Cross-Model Redaction Overlap with the GPT-5 Oracle

Dataset	Model	Overlap
ShareGPT	gpt-4.1-nano	0.203704
ShareGPT	gpt-4.1	0.256039
ShareGPT	claude-3-7-sonnet-20250219	0.182648
ShareGPT	claude-sonnet-4-20250514	0.178261
ShareGPT	lgai/exaone-deep-32b	0.090625
ShareGPT	mistralai/mistral-small-3.1-24b-instruct	0.189189
ShareGPT	qwen/qwen2.5-7b-instruct	0.152941
ShareGPT	qwen/qwen2.5-0.5b-instruct	0.014354
WildChat	gpt-4.1-nano	0.141414
WildChat	gpt-4.1	0.132653
WildChat	claude-3-7-sonnet-20250219	0.114035
WildChat	claude-sonnet-4-20250514	0.085470
WildChat	lgai/exaone-deep-32b	0.108974
WildChat	mistralai/mistral-small-3.1-24b-instruct	0.126214
WildChat	qwen/qwen2.5-7b-instruct	0.090909
WildChat	qwen/qwen2.5-0.5b-instruct	0.052326
MedQA	gpt-4.1-nano	0.023810
MedQA	gpt-4.1	0.093750
MedQA	claude-3-7-sonnet-20250219	0.060403
MedQA	claude-sonnet-4-20250514	0.175000
MedQA	lgai/exaone-deep-32b	0.031008
MedQA	mistralai/mistral-small-3.1-24b-instruct	0.111111
MedQA	qwen/qwen2.5-7b-instruct	0.090909
MedQA	qwen/qwen2.5-0.5b-instruct	0.026316
CaseHOLD	gpt-4.1-nano	0.000000
CaseHOLD	gpt-4.1	0.000000
CaseHOLD	claude-3-7-sonnet-20250219	0.200000
CaseHOLD	claude-sonnet-4-20250514	0.000000
CaseHOLD	lgai/exaone-deep-32b	0.000000
CaseHOLD	mistralai/mistral-small-3.1-24b-instruct	0.000000
CaseHOLD	qwen/qwen2.5-7b-instruct	0.000000
CaseHOLD	qwen/qwen2.5-0.5b-instruct	0.000000

Table 74: Cross-Model Abstraction Overlap with the GPT-5 Oracle

J Privacy Audit

J.1 Pooled Privacy Audit Across Oracle-Minimized Prompts - google/gemini-flash-1.5 & meta-llama/llama-3.1-70b-instruct as Attackers

action	N	p_{corr}	$p_{\text{corr,lo}}$	$p_{\text{corr,hi}}$	p_{unk}	$p_{\text{unk,lo}}$	$p_{\text{unk,hi}}$	$\overline{\text{conf}}$
abstract	679	0.119	0.097	0.146	0.323	0.288	0.359	0.627
redact	5627	0.077	0.070	0.084	0.762	0.750	0.773	0.175

Table 75: Span-wise recovery pooled across models by action on WildChat

action	N	p_{corr}	$p_{\text{corr,lo}}$	$p_{\text{corr,hi}}$	p_{unk}	$p_{\text{unk,lo}}$	$p_{\text{unk,hi}}$	$\overline{\text{conf}}$
abstract	1376	0.149	0.131	0.169	0.310	0.286	0.335	0.630
redact	6929	0.051	0.046	0.056	0.803	0.793	0.812	0.138

Table 76: Span-wise recovery pooled across models by action on ShareGPT

action	N	p_{corr}	$p_{\text{corr,lo}}$	$p_{\text{corr,hi}}$	p_{unk}	$p_{\text{unk,lo}}$	$p_{\text{unk,hi}}$	$\overline{\text{conf}}$
abstract	142	0.092	0.054	0.150	0.338	0.265	0.419	0.613
redact	6430	0.050	0.045	0.056	0.731	0.720	0.742	0.190

Table 77: Span-wise recovery pooled across models by action on CaseHOLD

action	N	p_{corr}	$p_{\text{corr,lo}}$	$p_{\text{corr,hi}}$	p_{unk}	$p_{\text{unk,lo}}$	$p_{\text{unk,hi}}$	$\overline{\text{conf}}$
abstract	935	0.056	0.043	0.072	0.030	0.021	0.043	0.964
redact	12835	0.027	0.024	0.030	0.790	0.783	0.797	0.158

Table 78: Span-wise recovery pooled across models by action on MedQA

Type	Hit@1 (orig)	Hit@1 (mask)	Hit@3 (orig)	Hit@3 (mask)
ADDRESS	1.000	0.000	1.000	0.000
AFFILIATION	0.681	0.017	0.729	0.021
AGE	1.000	0.000	1.000	0.000
ETHNICITY	0.000	0.000	0.000	0.000
FINANCIAL_INFORMATION	0.000	0.000	0.000	0.000
GEOLOCATION	0.760	0.148	0.792	0.167
HEALTH_INFORMATION	0.562	0.000	0.562	0.000
INCOME	1.000	0.000	1.000	0.000
NAME	0.918	0.000	0.999	0.000
RACE	0.735	0.000	0.735	0.000
TIME	0.870	0.006	0.916	0.006

Table 79: Type-wise recovery pooled by type on CaseHOLD

Type	N	H@1 CI	H@1 CI$^{\sim}$	H@3 CI	H@3 CI$^{\sim}$	$\overline{\text{conf}}$	$\overline{\text{conf}}^{\sim}$
ADDRESS	16	[80.6%, 100.0%]	[0.0%, 19.4%]	[80.6%, 100.0%]	[0.0%, 19.4%]	1.000000	1.000000
AFFILIATION	918	[65.0%, 71.0%]	[1.1%, 2.8%]	[69.9%, 75.7%]	[1.3%, 3.2%]	0.867000	0.799000
AGE	16	[80.6%, 100.0%]	[0.0%, 19.4%]	[80.6%, 100.0%]	[0.0%, 19.4%]	1.000000	0.000000
ETHNICITY	18	[0.0%, 17.6%]	[0.0%, 17.6%]	[0.0%, 17.6%]	[0.0%, 17.6%]	0.556000	0.100000
FINANCIAL_INFORMATION	18	[0.0%, 17.6%]	[0.0%, 17.6%]	[0.0%, 17.6%]	[0.0%, 17.6%]	1.000000	0.000000
GEOLOCATION	688	[72.7%, 79.1%]	[12.4%, 17.7%]	[76.0%, 82.1%]	[14.1%, 19.7%]	0.834000	0.521000
HEALTH_INFORMATION	16	[33.2%, 76.9%]	[0.0%, 19.4%]	[33.2%, 76.9%]	[0.0%, 19.4%]	1.000000	1.000000
INCOME	14	[78.5%, 100.0%]	[0.0%, 21.5%]	[78.5%, 100.0%]	[0.0%, 21.5%]	1.000000	0.557000
NAME	754	[89.6%, 93.5%]	[0.0%, 0.5%]	[99.3%, 100.0%]	[0.0%, 0.5%]	0.999000	0.608000
RACE	68	[62.0%, 82.6%]	[0.0%, 5.3%]	[62.0%, 82.6%]	[0.0%, 5.3%]	0.743000	0.500000
TIME	476	[83.7%, 89.7%]	[0.2%, 1.8%]	[88.8%, 93.8%]	[0.2%, 1.8%]	0.954000	0.635000

Table 80: Type-wise recovery on CaseHOLD: 95% confidence intervals (H@1/H@3; original and minimized) and mean top-1 confidence.

Type	Hit@1 (orig)	Hit@1 (mask)	Hit@3 (orig)	Hit@3 (mask)
AFFILIATION	0.000	0.000	0.000	0.000
AGE	0.992	0.000	0.996	0.000
DIETARY_PREFERENCE	0.125	0.000	0.125	0.000
GENDER	1.000	0.116	1.000	0.116
GEOLOCATION	0.705	0.000	0.705	0.000
HEALTH_INFORMATION	0.916	0.029	0.967	0.045
MARITAL STATUS	1.000	0.000	1.000	0.000
OCCUPATION	0.770	0.000	0.770	0.000
RACE	1.000	0.008	1.000	0.008
SEXUAL ORIENTATION	0.000	0.000	0.000	0.000
SEXUAL_ORIENTATION	0.517	0.000	0.517	0.000
TIME	0.533	0.000	0.822	0.000

Table 81: Type-wise recovery pooled by type on MedQA

Type	N	H@1 CI	H@1 CI$^{\sim}$	H@3 CI	H@3 CI$^{\sim}$	$\overline{\text{conf}}$	$\overline{\text{conf}}^{\sim}$
AFFILIATION	16	[0.0%, 19.4%]	[0.0%, 19.4%]	[0.0%, 19.4%]	[0.0%, 19.4%]	0.000000	0.000000
AGE	1530	[98.6%, 99.5%]	[0.0%, 0.3%]	[99.1%, 99.8%]	[0.0%, 0.3%]	1.000000	0.117000
DIETARY_PREFERENCE	16	[3.5%, 36.0%]	[0.0%, 19.4%]	[3.5%, 36.0%]	[0.0%, 19.4%]	1.000000	0.312000
GENDER	843	[99.5%, 100.0%]	[9.6%, 14.0%]	[99.5%, 100.0%]	[9.6%, 14.0%]	1.000000	0.826000
GEOLOCATION	190	[63.7%, 76.6%]	[0.0%, 2.0%]	[63.7%, 76.6%]	[0.0%, 2.0%]	0.705000	0.151000
HEALTH_INFORMATION	1424	[90.0%, 92.9%]	[2.2%, 4.0%]	[95.6%, 97.5%]	[3.5%, 5.7%]	1.000000	0.767000
MARITAL STATUS	16	[80.6%, 100.0%]	[0.0%, 19.4%]	[80.6%, 100.0%]	[0.0%, 19.4%]	1.000000	0.244000
OCCUPATION	122	[68.8%, 83.6%]	[0.0%, 3.1%]	[68.8%, 83.6%]	[0.0%, 3.1%]	0.885000	0.148000
RACE	121	[96.9%, 100.0%]	[0.1%, 4.5%]	[96.9%, 100.0%]	[0.1%, 4.5%]	1.000000	0.008000
SEXUAL ORIENTATION	14	[0.0%, 21.5%]	[0.0%, 21.5%]	[0.0%, 21.5%]	[0.0%, 21.5%]	1.000000	0.000000
SEXUAL_ORIENTATION	29	[34.4%, 68.6%]	[0.0%, 11.7%]	[34.4%, 68.6%]	[0.0%, 11.7%]	0.879000	0.817000
TIME	152	[45.4%, 61.0%]	[0.0%, 2.5%]	[75.4%, 87.5%]	[0.0%, 2.5%]	1.000000	0.203000

Table 82: Type-wise recovery on MedQA: 95% confidence intervals (H@1/H@3; original and minimized) and mean top-1 confidence.

Type	Hit@1 (orig)	Hit@1 (mask)	Hit@3 (orig)	Hit@3 (mask)
ADDRESS	1.000	0.000	1.000	0.000
AFFILIATION	0.845	0.042	0.892	0.044
AGE	0.746	0.029	0.787	0.029
EDUCATIONAL_RECORD	0.413	0.000	0.413	0.000
EMAIL	1.000	0.000	1.000	0.000
ETHNICITY	1.000	0.000	1.000	0.000
FINANCIAL_INFORMATION	0.833	0.148	0.852	0.148
GENDER	1.000	0.038	1.000	0.038
GEOLOCATION	0.858	0.051	0.961	0.058
HEALTH_INFORMATION	0.855	0.000	0.964	0.024
INCOME	0.729	0.000	0.729	0.000
IP_ADDRESS	0.429	0.000	1.000	0.000
MARITAL STATUS	0.655	0.000	0.745	0.000
MARITAL_STATUS	1.000	0.000	1.000	0.000
NAME	0.853	0.018	0.937	0.018
OCCUPATION	0.775	0.086	0.823	0.105
RACE	1.000	0.000	1.000	0.000
RELIGION	1.000	0.000	1.000	0.000
TIME	0.861	0.046	0.918	0.052
URL	0.922	0.000	0.933	0.000
VEHICLE	1.000	0.000	1.000	0.000

Table 83: Type-wise recovery pooled by type on ShareGPT

Type	N	H@1 CI	H@1 CI$^{\sim}$	H@3 CI	H@3 CI$^{\sim}$	$\overline{\text{conf}}$	$\overline{\text{conf}}^{\sim}$
ADDRESS	8	[67.6%, 100.0%]	[0.0%, 32.4%]	[67.6%, 100.0%]	[0.0%, 32.4%]	1.000000	0.125000
AFFILIATION	548	[81.2%, 87.3%]	[2.8%, 6.2%]	[86.4%, 91.6%]	[3.0%, 6.4%]	0.934000	0.703000
AGE	272	[69.1%, 79.4%]	[1.5%, 5.7%]	[73.4%, 83.1%]	[1.5%, 5.7%]	0.982000	0.293000
EDUCATIONAL_RECORD	46	[28.3%, 55.7%]	[0.0%, 7.7%]	[28.3%, 55.7%]	[0.0%, 7.7%]	0.900000	0.680000
EMAIL	17	[81.6%, 100.0%]	[0.0%, 18.4%]	[81.6%, 100.0%]	[0.0%, 18.4%]	1.000000	0.188000
ETHNICITY	31	[89.0%, 100.0%]	[0.0%, 11.0%]	[89.0%, 100.0%]	[0.0%, 11.0%]	1.000000	0.226000
FINANCIAL_INFORMATION	54	[71.3%, 91.0%]	[7.7%, 26.6%]	[73.4%, 92.3%]	[7.7%, 26.6%]	0.852000	0.907000
GENDER	78	[95.3%, 100.0%]	[1.3%, 10.7%]	[95.3%, 100.0%]	[1.3%, 10.7%]	1.000000	0.342000
GEOLOCATION	935	[83.4%, 87.9%]	[3.9%, 6.7%]	[94.7%, 97.2%]	[4.5%, 7.5%]	0.992000	0.674000
HEALTH_INFORMATION	83	[76.4%, 91.5%]	[0.0%, 4.4%]	[89.9%, 98.8%]	[0.7%, 8.4%]	1.000000	0.714000
INCOME	48	[59.0%, 83.4%]	[0.0%, 7.4%]	[59.0%, 83.4%]	[0.0%, 7.4%]	0.833000	0.677000
IP_ADDRESS	7	[15.8%, 75.0%]	[0.0%, 35.4%]	[64.6%, 100.0%]	[0.0%, 35.4%]	1.000000	0.857000
MARITAL STATUS	55	[52.3%, 76.6%]	[0.0%, 6.5%]	[61.7%, 84.2%]	[0.0%, 6.5%]	0.964000	0.251000
MARITAL_STATUS	9	[70.1%, 100.0%]	[0.0%, 29.9%]	[70.1%, 100.0%]	[0.0%, 29.9%]	1.000000	0.333000
NAME	621	[82.3%, 87.9%]	[1.0%, 3.1%]	[91.5%, 95.4%]	[1.0%, 3.1%]	0.958000	0.597000
OCCUPATION	209	[71.4%, 82.6%]	[5.5%, 13.2%]	[76.6%, 86.9%]	[7.1%, 15.4%]	0.911000	0.588000
RACE	13	[77.2%, 100.0%]	[0.0%, 22.8%]	[77.2%, 100.0%]	[0.0%, 22.8%]	1.000000	0.154000
RELIGION	7	[64.6%, 100.0%]	[0.0%, 35.4%]	[64.6%, 100.0%]	[0.0%, 35.4%]	1.000000	1.000000
TIME	656	[83.3%, 88.6%]	[3.2%, 6.5%]	[89.4%, 93.6%]	[3.7%, 7.2%]	0.998000	0.695000
URL	90	[84.8%, 96.2%]	[0.0%, 4.1%]	[86.2%, 96.9%]	[0.0%, 4.1%]	0.933000	0.561000
VEHICLE	6	[61.0%, 100.0%]	[0.0%, 39.0%]	[61.0%, 100.0%]	[0.0%, 39.0%]	1.000000	1.000000

Table 84: Type-wise recovery on ShareGPT: 95% confidence intervals (H@1/H@3; original and minimized) and mean top-1 confidence.

Type	Hit@1 (orig)	Hit@1 (mask)	Hit@3 (orig)	Hit@3 (mask)
AFFILIATION	0.830	0.019	0.871	0.019
AGE	0.691	0.000	0.764	0.000
EDUCATIONAL_RECORD	0.667	0.000	1.000	0.000
EMAIL	0.000	0.000	0.000	0.000
ETHNICITY	0.630	0.000	1.000	0.000
FINANCIAL_INFORMATION	0.923	0.000	0.923	0.000
GENDER	1.000	0.026	1.000	0.026
GEOLOCATION	0.898	0.022	0.954	0.031
GPA	1.000	0.000	1.000	0.000
HEALTH_INFORMATION	1.000	0.000	1.000	0.000
ID_NUMBER	1.000	0.000	1.000	0.000
INCOME	0.727	0.000	0.727	0.030
KEYS	1.000	0.000	1.000	0.000
NAME	0.903	0.000	0.981	0.000
OCCUPATION	0.854	0.080	0.934	0.080
PHONE_NUMBER	1.000	0.000	1.000	0.000
PRODUCT	1.000	0.000	1.000	0.000
QUANTITY	0.500	0.000	0.500	0.000
RACE	1.000	0.000	1.000	0.000
TIME	0.733	0.000	0.862	0.000
URL	0.886	0.000	0.886	0.000
USERNAME	0.533	0.000	0.533	0.000

Table 85: Type-wise recovery pooled by type on WildChat

Type	N	H@1 CI	H@1 CI$^{\sim}$	H@3 CI	H@3 CI$^{\sim}$	$\overline{\text{conf}}$	$\overline{\text{conf}}^{\sim}$
AFFILIATION	535	[79.6%, 85.9%]	[1.0%, 3.4%]	[84.0%, 89.7%]	[1.0%, 3.4%]	0.923000	0.671000
AGE	123	[60.5%, 76.6%]	[0.0%, 3.0%]	[68.2%, 83.1%]	[0.0%, 3.0%]	0.927000	0.263000
EDUCATIONAL_RECORD	24	[46.7%, 82.0%]	[0.0%, 13.8%]	[86.2%, 100.0%]	[0.0%, 13.8%]	1.000000	0.750000
EMAIL	18	[0.0%, 17.6%]	[0.0%, 17.6%]	[0.0%, 17.6%]	[0.0%, 17.6%]	0.200000	0.111000
ETHNICITY	27	[44.2%, 78.5%]	[0.0%, 12.5%]	[87.5%, 100.0%]	[0.0%, 12.5%]	1.000000	0.289000
FINANCIAL_INFORMATION	39	[79.7%, 97.3%]	[0.0%, 9.0%]	[79.7%, 97.3%]	[0.0%, 9.0%]	1.000000	0.949000
GENDER	38	[90.8%, 100.0%]	[0.5%, 13.5%]	[90.8%, 100.0%]	[0.5%, 13.5%]	1.000000	0.337000
GEOLOCATION	677	[87.3%, 91.9%]	[1.3%, 3.6%]	[93.6%, 96.8%]	[2.0%, 4.7%]	0.972000	0.577000
GPA	8	[67.6%, 100.0%]	[0.0%, 32.4%]	[67.6%, 100.0%]	[0.0%, 32.4%]	1.000000	0.250000
HEALTH_INFORMATION	39	[91.0%, 100.0%]	[0.0%, 9.0%]	[91.0%, 100.0%]	[0.0%, 9.0%]	1.000000	0.610000
ID_NUMBER	9	[70.1%, 100.0%]	[0.0%, 29.9%]	[70.1%, 100.0%]	[0.0%, 29.9%]	1.000000	0.111000
INCOME	33	[55.8%, 84.9%]	[0.0%, 10.4%]	[55.8%, 84.9%]	[0.5%, 15.3%]	0.758000	0.515000
KEYS	9	[70.1%, 100.0%]	[0.0%, 29.9%]	[70.1%, 100.0%]	[0.0%, 29.9%]	1.000000	0.444000
NAME	621	[87.8%, 92.4%]	[0.0%, 0.6%]	[96.7%, 98.9%]	[0.0%, 0.6%]	0.986000	0.558000
OCCUPATION	137	[78.5%, 90.3%]	[4.5%, 13.8%]	[88.0%, 96.5%]	[4.5%, 13.8%]	1.000000	0.531000
PHONE_NUMBER	9	[70.1%, 100.0%]	[0.0%, 29.9%]	[70.1%, 100.0%]	[0.0%, 29.9%]	1.000000	1.000000
PRODUCT	8	[67.6%, 100.0%]	[0.0%, 32.4%]	[67.6%, 100.0%]	[0.0%, 32.4%]	1.000000	0.750000
QUANTITY	18	[29.0%, 71.0%]	[0.0%, 17.6%]	[29.0%, 71.0%]	[0.0%, 17.6%]	1.000000	1.000000
RACE	8	[67.6%, 100.0%]	[0.0%, 32.4%]	[67.6%, 100.0%]	[0.0%, 32.4%]	1.000000	0.250000
TIME	536	[69.4%, 76.9%]	[0.0%, 0.7%]	[83.0%, 88.9%]	[0.0%, 0.7%]	0.998000	0.566000
URL	44	[76.0%, 95.0%]	[0.0%, 8.0%]	[76.0%, 95.0%]	[0.0%, 8.0%]	0.886000	0.443000
USERNAME	15	[30.1%, 75.2%]	[0.0%, 20.4%]	[30.1%, 75.2%]	[0.0%, 20.4%]	0.533000	0.533000

Table 86: Type-wise recovery on WildChat: 95% confidence intervals (H@1/H@3; original and minimized) and mean top-1 confidence.

J.2 GPT–5 as Attacker on Its Own Oracle-Minimized Prompts

action	N	p_{corr}	$p_{\text{corr,lo}}$	$p_{\text{corr,hi}}$	p_{unk}	$p_{\text{unk,lo}}$	$p_{\text{unk,hi}}$	$\overline{\text{conf}}$
abstract	679	0.119	0.097	0.146	0.323	0.288	0.359	0.627
redact	5627	0.077	0.070	0.084	0.762	0.750	0.773	0.175

Table 87: Span-wise recovery with GPT–5 as attacker on its own oracle-minimized prompts by action on WildChat

action	N	p_{corr}	$p_{\text{corr,lo}}$	$p_{\text{corr,hi}}$	p_{unk}	$p_{\text{unk,lo}}$	$p_{\text{unk,hi}}$	$\overline{\text{conf}}$
abstract	118	0.127	0.068	0.195	0.280	0.203	0.364	0.719
redact	987	0.020	0.012	0.029	0.967	0.954	0.978	0.030

Table 88: Span-wise recovery with GPT–5 as attacker on its own oracle-minimized prompts by action on ShareGPT

action	N	p_{corr}	$p_{\text{corr,lo}}$	$p_{\text{corr,hi}}$	p_{unk}	$p_{\text{unk,lo}}$	$p_{\text{unk,hi}}$	$\overline{\text{conf}}$
abstract	4	0.250	0.000	0.750	0.500	0.000	1.000	0.487
redact	397	0.055	0.035	0.081	0.922	0.894	0.947	0.069

Table 89: Span-wise recovery with GPT–5 as attacker on its own oracle-minimized prompts by action on CaseHOLD

action	N	p_{corr}	$p_{\text{corr,lo}}$	$p_{\text{corr,hi}}$	p_{unk}	$p_{\text{unk,lo}}$	$p_{\text{unk,hi}}$	$\overline{\text{conf}}$
abstract	21	0.000	0.000	0.000	0.000	0.000	0.000	1.000
redact	941	0.003	0.000	0.006	0.995	0.990	0.999	0.004

Table 90: Span-wise recovery with GPT–5 as attacker on its own oracle-minimized prompts by action on MedQA

Type	Hit@1 (orig)	Hit@1 (mask)	Hit@3 (orig)	Hit@3 (mask)
AFFILIATION	74.6%	0.0%	85.7%	0.0%
AGE	94.1%	0.0%	94.1%	0.0%
EDUCATIONAL_RECORD	66.7%	0.0%	66.7%	0.0%
EMAIL	0.0%	0.0%	0.0%	0.0%
ETHNICITY	50.0%	0.0%	75.0%	0.0%
FINANCIAL_INFORMATION	25.0%	0.0%	50.0%	0.0%
GENDER	100.0%	0.0%	100.0%	0.0%
GEOLOCATION	81.5%	0.0%	93.8%	0.0%
GPA	0.0%	0.0%	100.0%	0.0%
HEALTH_INFORMATION	20.0%	0.0%	40.0%	0.0%
ID_NUMBER	100.0%	0.0%	100.0%	0.0%
INCOME	50.0%	0.0%	50.0%	0.0%
KEYS	0.0%	0.0%	0.0%	0.0%
NAME	92.1%	0.0%	97.4%	0.0%
OCCUPATION	88.2%	0.0%	88.2%	0.0%
PHONE_NUMBER	100.0%	0.0%	100.0%	0.0%
PRODUCT	100.0%	0.0%	100.0%	0.0%
QUANTITY	0.0%	0.0%	50.0%	0.0%
RACE	100.0%	0.0%	100.0%	0.0%
TIME	73.0%	0.0%	82.5%	0.0%
URL	60.0%	0.0%	60.0%	0.0%
USERNAME	50.0%	0.0%	50.0%	0.0%

Table 91: Type-wise recovery with GPT–5 as attacker on its own oracle-minimized prompts on WildChat.

Type	Hit@1 (orig)	Hit@1 (mask)	Hit@3 (orig)	Hit@3 (mask)
ADDRESS	100.0%	0.0%	100.0%	0.0%
AFFILIATION	74.2%	1.5%	81.8%	1.5%
AGE	58.3%	0.0%	63.9%	0.0%
EDUCATIONAL_RECORD	33.3%	0.0%	33.3%	0.0%
EMAIL	100.0%	0.0%	100.0%	0.0%
ETHNICITY	75.0%	0.0%	75.0%	0.0%
FINANCIAL_INFORMATION	42.9%	0.0%	57.1%	0.0%
GENDER	90.9%	0.0%	100.0%	0.0%
GEOLOCATION	82.2%	1.7%	93.2%	1.7%
HEALTH_INFORMATION	50.0%	0.0%	60.0%	0.0%
INCOME	33.3%	0.0%	50.0%	0.0%
IP_ADDRESS	100.0%	0.0%	100.0%	0.0%
MARITAL_STATUS	37.5%	0.0%	50.0%	0.0%
MARITAL_STATUS	0.0%	0.0%	0.0%	0.0%
NAME	91.0%	1.3%	94.9%	1.3%
OCCUPATION	59.3%	3.7%	70.4%	3.7%
RACE	100.0%	0.0%	100.0%	0.0%
RELIGION	100.0%	0.0%	100.0%	0.0%
TIME	71.4%	2.4%	83.3%	2.4%
URL	90.9%	0.0%	90.9%	0.0%
VEHICLE	100.0%	0.0%	100.0%	0.0%

Table 92: Type-wise recovery with GPT–5 as attacker on its own oracle-minimized prompts on ShareGPT.

Type	Hit@1 (orig)	Hit@1 (mask)	Hit@3 (orig)	Hit@3 (mask)
ADDRESS	100.0%	0.0%	100.0%	0.0%
AFFILIATION	74.5%	1.8%	81.8%	1.8%
AGE	0.0%	0.0%	0.0%	0.0%
ETHNICITY	0.0%	0.0%	0.0%	0.0%
FINANCIAL_INFORMATION	100.0%	0.0%	100.0%	0.0%
GEOLOCATION	83.3%	0.0%	95.2%	2.4%
HEALTH_INFORMATION	0.0%	0.0%	0.0%	0.0%
INCOME	0.0%	0.0%	100.0%	0.0%
NAME	84.4%	0.0%	88.9%	0.0%
RACE	50.0%	0.0%	50.0%	0.0%
TIME	82.8%	0.0%	93.1%	0.0%

Table 93: Type-wise recovery with GPT–5 as attacker on its own oracle-minimized prompts on CaseHOLD.

Type	Hit@1 (orig)	Hit@1 (mask)	Hit@3 (orig)	Hit@3 (mask)
AFFILIATION	0.0%	0.0%	0.0%	0.0%
AGE	99.0%	0.0%	99.0%	0.0%
DIETARY_PREFERENCE	100.0%	0.0%	100.0%	0.0%
GENDER	100.0%	0.0%	100.0%	0.0%
GEOLOCATION	46.2%	0.0%	53.8%	0.0%
HEALTH_INFORMATION	83.7%	0.0%	92.4%	0.0%
MARITAL_STATUS	100.0%	0.0%	100.0%	0.0%
OCCUPATION	50.0%	0.0%	50.0%	0.0%
RACE	100.0%	0.0%	100.0%	0.0%
SEXUAL_ORIENTATION	100.0%	0.0%	100.0%	0.0%
TIME	50.0%	0.0%	80.0%	0.0%

Table 94: Type-wise recovery with GPT–5 as attacker on its own oracle-minimized prompts on MedQA.