Mind the Implicit Gap: Quantifying Attribute Inference Leakage in Privacy-Conscious Delegation

Nicole Ren*
GovTech Singapore
nicole_ren@tech.gov.sg

James Teo*
GovTech Singapore
james_teo@tech.gov.sg

Abstract

Even when explicit Personally Identifiable Information (PII) is removed from LLM queries, implicit contextual cues can still reveal sensitive user attributes. Yet current Privacy-Conscious Delegation frameworks overlook this threat entirely, and existing PII span-based evaluation metrics cannot address it. We address this gap on both fronts. First, we introduce Attribute Inference Leakage (AIL), a label-free metric that quantifies implicit privacy leakage through set-based semantic comparison of attributes inferred before and after sanitisation. Using AIL, we provide the first empirical evidence in Privacy-Conscious Delegation that even perfect PII removal eliminates only 18% of implicit leakage relative to direct delegation, leaking the vast majority of inferable attributes. Second, we propose GIST, a training-free approach for context-aware generalisation to protect both explicit and implicit privacy while preserving response quality. Experiments across six local models (1B to 8B parameters) show GIST consistently achieves a Pareto improvement over state-of-the-art PAPILLON on all three evaluation dimensions of quality, explicit and implicit leakage. GIST also consistently reduces implicit leakage below the explicit leakage lower bound (perfect PII removal), hence GIST addresses a privacy dimension that PII-centric methods cannot.

1 Introduction

The query "What about my husband?" contains no Personally Identifiable Information (PII) yet reveals the user's relationship status. LLMs can now infer such private attributes from seemingly innocuous text with near-human accuracy (Staab et al., 2024), creating a scalable privacy threat. As users routinely share sensitive details with untrusted LLMs (Mireshghallah et al., 2024), this implicit leakage poses a pressing concern.

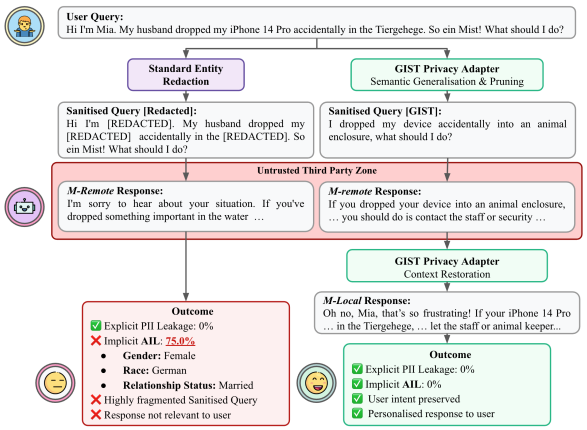


Figure 1: Standard redaction (left) leaks implicit attributes and degrades response quality. GIST (right) generalises task-relevant details ("Tiergehege" → "animal enclosure") and neutralises profiling cues, simultaneously preserving response quality and defending against attribute inference attacks (reduce AIL).

Privacy-Conscious Delegation addresses inference-time privacy risks by chaining a trusted local model with an untrusted remote model, enabling users to leverage proprietary LLMs while protecting sensitive information (Siyan et al., 2025). However, existing frameworks focus exclusively on explicit PII (Siyan et al., 2025; Hui et al., 2026), leaving implicit contextual cues that fuel attribute inference entirely unaddressed. This blind spot extends to evaluation: standard span-based metrics that detect PII cannot address implicit leakage, and recent adversarial accuracy metrics require ground-truth attribute labels (Staab et al., 2025) that are rarely available in real-world scenarios.

We address this gap on both fronts. First, we introduce **Attribute Inference Leakage (AIL)**, a label-free metric that quantifies implicit privacy leakage by comparing attribute inferability before and after sanitisation with a set-based semantic comparison. Second, we propose **GIST**

*Equal contribution

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 21080–21114
October 24-29, 2026 ©2026 Association for Computational Linguistics

(**G**eneralising **I**dentity for **S**anitising **T**ext), a training-free approach for context-aware generalisation, transforming task-relevant sensitive details into abstractions while pruning non-essential profiling cues hence preserving the "gist" of user intent while preventing attribute inference (Figure 1). We make the following contributions:

- We formalise implicit attribute inference as a privacy dimension for Privacy-Conscious Delegation and introduce **AIL**, a label-free metric to quantify implicit leakage through set-based semantic comparison of inferred attributes.
- Using AIL, we provide the first empirical evidence in Privacy-Conscious Delegation that even perfect PII removal eliminates only 18% of implicit leakage relative to direct delegation, exposing the vast majority of inferable attributes. Current PII-centric evaluation critically misses this residual privacy risk.
- We propose **GIST**, a training-free approach for context-aware generalisation to protect both explicit and implicit privacy while preserving response quality. GIST consistently achieves a Pareto improvement over the state-of-the-art PAPILLON on all three metrics across six local models (1B to 8B parameters), and consistently reduces implicit leakage below the explicit leakage lower bound (perfect PII removal), hence GIST addresses a privacy dimension that PII-centric methods cannot.
- We release an **enhanced PUPA-TNB benchmark**[1] to correct PII annotations, establishing a more reliable testbed for further research.

2 Related Work

Privacy Risks in LLM Inference. The increasing use of LLMs hosted by untrusted external providers has introduced significant privacy risks. Beyond training-time risks due to memorisation and leakage of sensitive data (Carlini et al., 2023), inference-time privacy poses critical challenges as users routinely share sensitive personal information including financial details with LLM applications (Mireshghallah et al., 2024).

[1]The dataset is available at https://github.com/njemnlp2026/pupa-tnb-enhanced.

Privacy-Conscious Delegation. To address inference-time privacy risks, Privacy-Conscious Delegation has emerged to chain trusted local models with untrusted remote models to leverage superior capabilities of third-party LLMs while protecting privacy (Siyan et al., 2025). PAPILLON (Siyan et al., 2025) pioneered this with query rewriting to remove PII in queries. Privacy-R1 (Hui et al., 2026) extends this paradigm through reinforcement learning to learn a policy that shields non-essential PII while deliberately exposing task-critical PII for utility. However, the exclusive focus on explicit PII protection in existing methods creates a critical gap by failing to address residual implicit contextual cues that enable attribute inference attacks. This is particularly concerning as sensitive personal information commonly shared in human-chatbot interactions such as job applications remains embedded despite PII removal (Mireshghallah et al., 2024). Our work addresses this gap by introducing both a label-free metric (AIL) to quantify implicit leakage and a training-free generalisation approach (GIST) that protects against both explicit and implicit privacy risks without sacrificing utility.

Attribute Inference Attacks. Recent work demonstrated that LLMs can predict personal attributes from seemingly non-revealing online posts with near human-level accuracy (Staab et al., 2024). This highlights a severe and scalable privacy threat that current delegation frameworks remain vulnerable to as they lack mechanisms in both measuring and mitigating this implicit privacy leakage.

Evaluation Metrics for Privacy Protection. Existing evaluation frameworks predominantly rely on span-based metrics such as precision and recall that measure detection of explicit PII. Although useful for traditional PII protection, they are insufficient to capture the privacy risks of contextual information leakage. Recent work introduced adversarial accuracy to measure how accurately adversarial LLMs can infer personal attributes from sanitised text given human-annotated attribute labels (Staab et al., 2025). However, this requires ground truth annotations that are rarely available due to the high cost of manual annotation.

The existing landscape highlights a critical gap: while current Privacy-Conscious Delegation frameworks address the privacy risks from explicit PII, they remain vulnerable to the emerging privacy threat of scalable attribute inference attacks. We

address this by introducing AIL, a novel label-free metric to quantify implicit information leakage, and GIST, a mitigation approach leveraging context-aware generalisation for comprehensive privacy protection while preserving user intent for utility.

3 Problem Formulation

3.1 Task Setting

Privacy-Conscious Delegation (Siyan et al., 2025) operates with a trusted local model ($\mathcal{M}_{\text{local}}$) deployed in the user's secure environment and an untrusted remote model ($\mathcal{M}_{\text{remote}}$) accessed via API. Given a query q containing PII units $P = \{p_1, p_2, \ldots, p_k\}$, the existing objective is to synthesise a prompt q' that enables $\mathcal{M}_{\text{remote}}$ to produce a response r comparable to the target r^* (obtained from using q directly) while minimising PII exposure. This formulation focuses on single-turn sessions. Evaluation uses quality preservation and PII leakage, which are LLM-as-a-Judge (Zheng et al., 2023) metrics validated in prior work (Siyan et al., 2025).

3.2 Extended Objective

The existing objective addresses only explicit PII for privacy protection. However, queries without PII can still leak sensitive attributes through implicit contextual cues. For example, "What about my husband?" contains no PII yet reveals the user's relationship status (married). We therefore extend the privacy considerations beyond explicit PII exposure to encompass implicit leakage, where an adversary leverages implicit contextual cues to deduce sensitive attributes from user queries. The synthesised prompt q' exposed to $\mathcal{M}_{\text{remote}}$ now needs to satisfy three competing objectives: (1) **Quality Preservation**: q' retains sufficient semantic fidelity for $\mathcal{M}_{\text{remote}}$ to generate a response r comparable to r^*; (2) **Explicit Privacy Protection**: q' minimises exposure of identifiable PII units; and (3) **Implicit Privacy Protection**: q' limits the potential for inferring sensitive personal attributes from residual contextual cues. The core challenge lies in effectively balancing these objectives: maximising response quality while simultaneously mitigating both explicit and implicit information leakage.

4 Attribute Inference Leakage

The existing privacy evaluation metric for Privacy-Conscious Delegation detects explicit PII exposed to $\mathcal{M}_{\text{remote}}$. However, this overlooks how residual implicit cues in the synthesised prompt q' may still enable adversaries to infer sensitive user attributes even after PII removal. We introduce Attribute Inference Leakage (AIL), a novel metric that quantifies the degree to which personal attributes remain inferable from sanitised text compared to original input (Figure 2). Unlike recent adversarial accuracy that requires labels (Staab et al., 2025), AIL simulates attribute inference attacks on both original query q and synthesised query q' to measure the change in inferability using LLM-as-a-Judge (Zheng et al., 2023). By focusing on the relative change in inferability rather than absolute attribute detection, AIL eliminates the need for ground truth labels, making it practical for resource-constrained scenarios while providing more comprehensive assessment of privacy leakage that existing PII-based metrics miss. All prompts are in Appendix A.

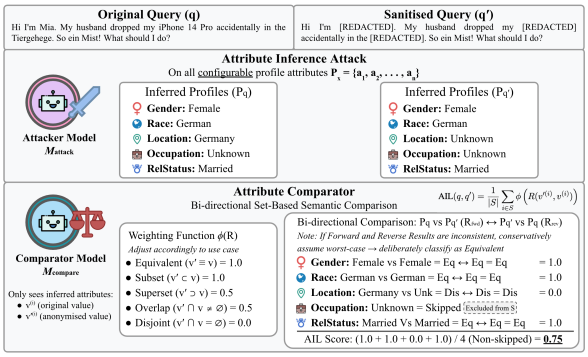


Figure 2: Overview of AIL. An attacker model infers attributes from both original and sanitised queries. A separate comparator model performs bi-directional set-based semantic comparison of corresponding attribute pairs. Risk scores are then averaged over all originally inferable attributes to quantify residual implicit leakage.

4.1 Simulated Inference Attack

Given the demonstrated capability of LLMs to perform human-level attribute inferences (Staab et al., 2024), we use an attacker model, $\mathcal{M}_{\text{attack}}$, to simulate adversarial attribute inference attacks via Chain-of-Thought (Wei et al., 2022). For each input text x, $\mathcal{M}_{\text{attack}}$ extracts a structured attribute profile $P_x = \{a_1, a_2, \ldots, a_n\}$ covering five demographic attributes: *Gender, Race, Location, Occupation,* and *Relationship Status*. We perform this extraction on both q and q':

$$P_q = \mathcal{M}_{\text{attack}}(q), \quad P_{q'} = \mathcal{M}_{\text{attack}}(q') \quad (1)$$

To ensure comparability, $\mathcal{M}_{\text{attack}}$ is constrained to produce categorical outputs for closed-set attributes (e.g., *Gender, Relationship Status*) and standardised descriptions for open-set attributes. $\mathcal{M}_{\text{attack}}$ is instructed to avoid guesses based on stereotypes and biases, labelling attributes as `Unknown` when unable to reliably infer from the available context.

4.2 Set-Based Semantic Comparison

Standard string match is inadequate for comparing open-ended attributes (e.g., "Developer" vs "Engineer"), while numerical scales for LLM-as-a-Judge can result in inconsistencies (Li et al., 2026b). Thus, we introduce set-based pairwise comparison, treating each extracted attribute as a semantic set of entities satisfying the description. For each attribute i, we employ a separate LLM comparator $\mathcal{M}_{\text{compare}}$ to determine the set relation R between sanitised value $v'^{(i)}$ and original value $v^{(i)}$. $\mathcal{M}_{\text{compare}}$ is distinct from $\mathcal{M}_{\text{attack}}$ by design to eliminate circularity where a model evaluates its own output. $\mathcal{M}_{\text{compare}}$ receives only $(v'^{(i)}, v^{(i)})$, isolated from the source text and $\mathcal{M}_{\text{attack}}$'s reasoning. This strict information barrier ensures $\mathcal{M}_{\text{compare}}$ performs a purely semantic comparison. We define five distinct relationships, ordered by severity:

- **Equivalent ($v' \equiv v$):** Identical semantic scope, includes synonyms and variations (e.g., "us" vs "usa"). This denotes a privacy failure.
- **Subset ($v' \subset v$):** Privacy leakage through increased specificity, sanitised value is more specific than the original (e.g., v = "Asian", v' = "Korean").
- **Superset ($v' \supset v$):** Partial privacy preservation, sanitised value generalises the original (e.g., v' = "France", v = "Paris").
- **Overlap ($v' \cap v \neq \emptyset$):** The values intersect but neither contains the other (e.g., "Researcher" vs "Professor"). This represents partial privacy preservation with semantic ambiguity.
- **Disjoint ($v' \cap v = \emptyset$):** The values share no commonality, or the attribute becomes `Unknown` ($\perp$). This represents effective obfuscation and optimal privacy preservation.

To mitigate positional bias in LLM pairwise judgements (Zheng et al., 2023), we enforce **bi-directional consistency** by computing the comparison in both forward (v' relative to v) and reverse (v relative to v') directions. A comparison is only valid when the logical converse holds (e.g., if R_{fwd} is Superset, then R_{rev} must be Subset). When the comparison is invalid, it defaults to Equivalent as a conservative risk penalty.

4.3 AIL Scoring

AIL quantifies residual privacy risk with a weighting function that maps set relations to risk scores:

$$\phi(R) = \begin{cases} w_{\text{eq}} & \text{if } R = \text{Equivalent} \\ w_{\text{sub}} & \text{if } R = \text{Subset} \\ w_{\text{sup}} & \text{if } R = \text{Superset} \\ w_{\text{ovl}} & \text{if } R = \text{Overlap} \\ 0.0 & \text{if } R = \text{Disjoint} \end{cases} \tag{2}$$

subject to the constraint:

$$w_{\text{eq}} \geq w_{\text{sub}} \geq w_{\text{sup}} \geq w_{\text{ovl}} \geq 0 \tag{3}$$

This ensures relationship types with greater privacy risk receive higher weights. The design parallels the F_β-score's β parameter: just as β controls the relative importance of precision and recall, AIL's weights allow practitioners to calibrate the importance of different types of partial leakage. We adopt **default conservative weights**: $w_{\text{eq}} = w_{\text{sub}} = 1.0$ and $w_{\text{sup}} = w_{\text{ovl}} = 0.5$. Under this configuration, Equivalent and Subset (where the adversary infers at least as much) incur full penalty, while Superset and Overlap (partial information) incur half penalty. The overall AIL score is the average risk across all attributes originally inferable, denoted by the index set S:

$$\text{AIL}(q, q') = \frac{1}{|S|} \sum_{i \in S} \phi\left(R(v'^{(i)}, v^{(i)})\right) \tag{4}$$

Attributes that are `Unknown` in q are excluded from S, as AIL measures protection of *initially inferable* information rather than penalising noise introduced via sanitisation. A score of 0.0 represents perfect obfuscation; 1.0 represents complete privacy failure. See algorithm in Appendix A.

5 GIST: A Mitigation Approach

Existing Privacy-Conscious Delegation methods focus on explicit PII removal but miss implicit contextual cues that still enable attribute inference attacks. Thus, we introduce **GIST** (**G**eneralising **I**dentity for **S**anitising **T**ext) which addresses both

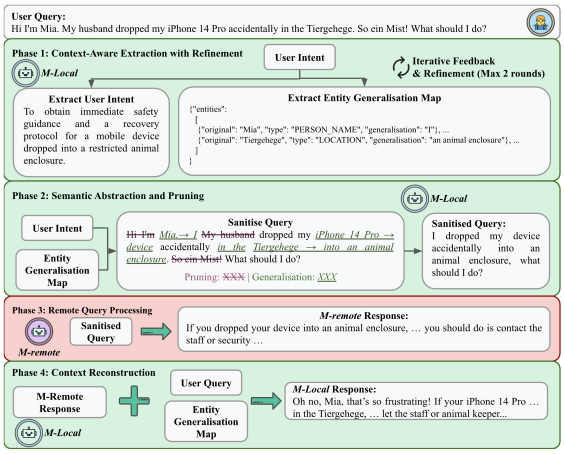


Figure 3: Overview of GIST: (1) Context-Aware Extraction of user intent and entities, (2) Semantic Abstraction & Pruning generalises task-relevant entities and prunes non-essential ones, (3) Inference by untrusted $\mathcal{M}_{\text{remote}}$, and (4) Context Restoration for final response.

explicit and implicit privacy risks while maintaining response quality. GIST leverages a small language model $\mathcal{M}_{\text{local}}$ to perform context-aware semantic abstraction by generalising task-relevant sensitive details while pruning non-essential identifying cues. Unlike Privacy-R1 (Hui et al., 2026), GIST is training-free and avoids exposing task-critical PII by generalising rather than selective exposure. The pipeline operates in four phases (Figure 3). All prompts are in Appendix B.

5.1 Phase 1: Context-Aware Extraction

$\mathcal{M}_{\text{local}}$ extracts the user's core intent $\mathcal{I}$ and a protection set $\mathcal{E}$ of sensitive spans based on nine classes (*PersonName*, *Organisation*, *Location*, *Date*, *NationalityReligionPolitical*, *Financial*, *Occupation*, *URL*, and *UniqueIdentifier*) via Chain-of-Thought (Wei et al., 2022) and few-shot (Brown et al., 2020) prompting with constrained decoding to enforce adherence to JSON schemas (Li et al., 2026a). Each entity $e_i \in \mathcal{E}$ is paired with a context-aware generalisation (e.g., "Mark" $\rightarrow$ "colleague A") that preserves narrative roles without revealing identity. To mitigate the inadequacy of single-pass extraction, a separate validator instance implements a self-refinement loop (Madaan et al., 2023): it checks for residual entities and verifies entities exist within the original query. Upon detecting violations, the validator generates structured feedback and the extraction regenerates; this repeats for a default maximum of k=2 iterations (see Table 13) to balance extraction robustness with speed.

5.2 Phase 2: Semantic Abstraction & Pruning

Given $(\mathcal{I}, \mathcal{E})$, $\mathcal{M}_{\text{local}}$ generates the sanitised $q' = \text{Abstract}(q, \mathcal{I}, \mathcal{E})$ through the following operations:

Context-Aware Generalisation. $\mathcal{M}_{\text{local}}$ contextually replaces all entity spans and their morphological variants (pronouns, possessives, inflected forms) with the generalisations from $\mathcal{E}$. Unlike rigid string replacement, this retains narrative coherence (e.g., "my daughter Sarah" $\rightarrow$ "a family member" rather than "my daughter [PERSON]").

Perspective Neutralisation and Pruning. To mitigate implicit demographic inference, $\mathcal{M}_{\text{local}}$ transforms first-person narratives into objective third-person descriptions, eliminating authorial voice markers including emotional expressions and dialectal features that may leak demographic attributes like age, gender or cultural background. Non-essential identifying details are concurrently pruned to minimise the attack surface.

The resulting sanitised prompt q' preserves the semantic "gist" of the user's intent while removing both explicit and implicit sensitive contextual cues.

5.3 Phases 3 & 4: Remote Inference & Context Restoration

The sanitised q' is sent to $\mathcal{M}_{\text{remote}}$ for inference $(r' = \mathcal{M}_{\text{remote}}(q'))$. As GIST uses contextually meaningful abstractions rather than destructive masks, it retains sufficient context for high-quality responses. Finally, $\mathcal{M}_{\text{local}}$ restores the user's original entities into r' via $r = \text{Restore}(q, \mathcal{E}, r')$, resolving morphological dependencies and grammatical agreement (e.g., "the person submitted their work" $\rightarrow$ "John submitted his work").

6 Experimental Setup

6.1 Dataset

We evaluate on our enhanced version of the PUPA-TNB dataset (Siyan et al., 2025) designed for Privacy-Conscious Delegation tasks. PUPA-TNB is derived from real-world user-LLM chats in the WildChat dataset (Zhao et al., 2024) and contains realistic PII across diverse domains. Each instance contains a user query, target response, and annotated PII units.

6.1.1 Enhancement to Dataset

Upon reviewing PUPA-TNB, we surfaced several issues that could compromise benchmarks. We

found 28 Presidio (Mendels et al., 2018) placeholders in user queries with missing PII units, creating disparities with their corresponding target responses. Moreover, 25 records exhibited incorrect PII units that either referenced absent entities in the user query or provided incomplete coverage of the PIIs present. One record lacked a target response entirely. These anomalies would hinder evaluations that use both target responses and PII units.

To address these limitations, we created an enhanced version of PUPA-TNB. First, we generated synthetic replacements for 28 Presidio placeholders in user queries using Llama-3.3-70B-Instruct (Llama Team, AI @ Meta, 2024), ensuring that the queries contained realistic personal information. Second, we manually corrected PII unit annotations in 25 records to ensure consistency with the queries. Third, we regenerated target responses using GPT-4 (OpenAI, 2023) for the missing record and for all queries that received synthetic PII enhancements. We used GPT-4 for response generation to maintain consistency with the original PUPA-TNB dataset construction methodology (Siyan et al., 2025).

This enhanced dataset enables more meaningful evaluation of privacy-preserving approaches by providing more complete and consistent annotations (see examples in Appendix C). Given the nascent state of research in Privacy-Conscious Delegation, we contribute this refined dataset to facilitate further research.

6.2 Baselines

We establish baselines across the privacy-utility spectrum. **Direct Delegation** to $\mathcal{M}_{\text{remote}}$ sets the theoretical utility upper bound. **PII Redaction** of ground-truth PII units sets the theoretical lower bound of explicit privacy leakage. **PAPILLON** represents state-of-the-art in Privacy-Conscious Delegation using zero-shot query rewriting to remove explicit PII (Siyan et al., 2025). While unreleased code for Privacy-R1 (Hui et al., 2026) at the time of writing prevented reproduction, GIST maintains theoretical benefits. Privacy-R1 deliberately exposes task-critical PII for utility and introduces stateful complexity, whereas GIST is a training-free framework that generalises task-critical PII and mitigates implicit privacy risks in addition to PII protection.

6.3 Models

For reproducibility, we use open-weight models with temperature $= 0$ and top-p $= 1.0$ for all inferences. For $\mathcal{M}_{\text{local}}$, we use instruction-tuned models across diverse parameter sizes: `gemma-3-1b-it` (Gemma Team, Google DeepMind, 2025), `Llama-3.2-3B-Instruct` (Llama Team, AI @ Meta, 2024), `Phi-4-mini-instruct` (Microsoft, 2025), `gemma-3-4b-it` (Gemma Team, Google DeepMind, 2025), `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023) and `Llama-3.1-8B-Instruct` (Llama Team, AI @ Meta, 2024). All $\mathcal{M}_{\text{local}}$ inferences are executed via the vLLM engine (Kwon et al., 2023) on a single NVIDIA L4 (24 GB VRAM). We enforce structured JSON outputs with constrained decoding using XGrammar (Li et al., 2026a). `gpt-oss-120b` (OpenAI, 2025) is used as $\mathcal{M}_{\text{remote}}$. For evaluation, `Kimi-K2.5` (Kimi Team, 2026) serves as the judge for Quality Preservation, Privacy Leakage and $\mathcal{M}_{\text{compare}}$ for AIL, with `GLM-4.7` (GLM Team, 2025) as $\mathcal{M}_{\text{attack}}$.

6.4 Evaluation Metrics

We use the two existing metrics validated in prior work (Siyan et al., 2025). **Quality Preservation** (%) measures the percentage of responses judged equivalent or better than target response. **Privacy Leakage** (%) quantifies the average fraction of PII units exposed to the remote model. Additionally, we use **AIL** (%) to assess implicit privacy leakage.

6.5 Validation of AIL

We validated AIL with unpaid human annotators (non-author) with experience in using anonymisation tools (instructions in Appendix D). We measured inter-annotator reliability via Krippendorff's α and majority-vote human-LLM alignment via Percentage Agreement (PA). Given the open-ended nature of attribute inference, five annotators validated 50 queries against LLM-inferred profiles, yielding high consensus ($PA = 92\%$, $\alpha = 0.81$). This confirms that the LLM reliably detects implicit cues consistent with human interpretation, aligning with prior work (Staab et al., 2024). For set-based comparison, three annotators labelled 150 attribute pairs balanced across the five types, achieving strong overall alignment ($PA = 95.6\%$). Agreement was unanimous for all types except Overlap ($PA = 77.8\%$, $\alpha = 0.82$), reflecting inherent subjectivity in partial semantic intersections.

Model / Approach	Quality (%)	$\Delta_{Quality}$	Leakage (%)	$\Delta_{Leakage}$	AIL (%)	Δ_{AIL}
gpt-oss-120b						
Direct Delegation	97.1	–	100.0	–	52.1	–
PII Redaction	74.0	–	0.0	–	42.8	–
gemma-3-1b-it						
PAPILLON	17.2	–	43.7	–	40.8	–
GIST	23.3	+6.1	22.4	-21.3	**21.8**	**-19.0**
Llama-3.2-3B-Instruct						
PAPILLON	34.2	–	28.7	–	37.8	–
GIST	37.3	+3.1	7.0	-21.7	32.4	-5.4
Phi-4-mini-instruct						
PAPILLON	54.0	–	46.0	–	39.9	–
GIST	58.0	+4.0	8.8	**-37.2**	30.3	-9.6
gemma-3-4b-it						
PAPILLON	52.3	–	14.1	–	33.9	–
GIST	**66.7**	**+14.4**	**1.6**	-12.5	27.6	-6.3
Mistral-7B-Instruct-v0.3						
PAPILLON	48.1	–	8.4	–	35.0	–
GIST	55.9	+7.8	5.0	-3.4	27.6	-7.4
Llama-3.1-8B-Instruct						
PAPILLON	51.7	–	8.4	–	35.2	–
GIST	57.4	+5.7	5.9	-2.5	29.5	-5.7

Table 1: Comparison of GIST to baselines, gpt-oss-120b as remote model and six local models on enhanced PUPA-TNB. GIST consistently outperforms PAPILLON across all local models and metrics. Best Privacy-Conscious Delegation result for each metric and biggest differences between GIST and PAPILLON are bolded.

7 Results and Analysis

7.1 AIL Diagnostic Findings

AIL reveals a fundamental blind spot in existing privacy evaluation in Privacy-Conscious Delegation. PII Redaction, theoretical lower bound for explicit leakage (0%), still exhibits 42.8% AIL, only 18% relative reduction from Direct Delegation (Table 1). This is, to our knowledge, the first empirical evidence in Privacy-Conscious Delegation that perfect PII removal leaves substantial implicit leakage.

Despite query rewriting, PAPILLON's AIL ranges from 33.9% (gemma-3-4b-it) to 40.8% (gemma-3-1b-it) in Table 1, confirming that its exclusive focus on explicit PII removal does not address the implicit attack surface. Together, these results demonstrate that (i) **PII removal alone does not address attribute inference risks**, and (ii) **current evaluation frameworks for Privacy-Conscious Delegation, which report only explicit leakage, consequently miss this residual privacy risk**.

7.2 GIST Mitigation Results

Table 1 presents the main results across six local models (1B to 8B parameters), GIST consistently outperforms PAPILLON on all metrics.

Pareto Improvement over PAPILLON. GIST simultaneously improves quality and strengthens privacy, representing a Pareto improvement. Quality rises by 3.1 to 14.4 pp, while explicit Leakage drops by up to 37.2 pp. Notably, gemma-3-4b-it with GIST achieves 1.6% Leakage, the closest to PII Redaction's theoretical bound (0%) while delivering 66.7% Quality versus PAPILLON's 52.3%, closing the gap to Direct Delegation with far stronger privacy protection. We attribute this to GIST's use of contextually meaningful generalisations rather than destructive redaction or rewriting, such that the remote model has sufficient semantic context for high-quality responses while identifying specifics are abstracted away.

Implicit Leakage Protection Beyond PII Removal. GIST is the first Privacy-Conscious Delegation approach to demonstrably reduce implicit attribute inference leakage across all six local models tested. With gemma-3-1b-it, GIST achieves 21.8% AIL, nearly half of both PII Redaction (42.8%) and PAPILLON (40.8%) despite using only a 1B-parameter model. This confirms that perspective neutralisation and context-aware pruning in GIST address a threat vector that PII-centric methods cannot mitigate regardless of their explicit protection efficacy, and that GIST's structured pipeline compensates for limited model capability.

Qualitative Analysis. Qualitative inspection reveals two broad categories of implicit cues that

Attacker / Comparator	Δ AIL (GIST − PAPILLON) ↓ gemma-3 1b	Phi-4 mini	Mistral 7B
Opus-4.6			
Kimi-K2.5	-23.5	-19.3	-6.2
GLM-4.7	-21.9	-16.8	-5.7
Avg ± SD	-22.7 ± 1.1	-18.1 ± 1.8	-6.0 ± 0.4
Kimi-K2.5			
Opus-4.6	-14.1	-12.9	-6.1
GLM-4.7	-16.7	-11.4	-5.6
Avg ± SD	-15.4 ± 1.8	-12.2 ± 1.1	-5.9 ± 0.4
GLM-4.7			
Opus-4.6	-21.8	-10.5	-8.3
Kimi-K2.5	-19.0	-9.6	-7.4
Avg ± SD	-20.4 ± 2.0	-10.1 ± 0.6	-7.9 ± 0.6

Table 2: AIL robustness across Attacker × Comparator pairs. Δ AIL = GIST − PAPILLON (%); negative values indicate reduced implicit leakage. All values are negative and within-attacker standard deviations remain low (0.4 to 2.0 pp), confirming that AIL's relative rankings are stable regardless of LLM configuration.

AIL captures and PII-centric methods miss (see examples in Appendix E): (1) *domain-specific markers*: terminology and cultural references that enable inference (e.g., location, occupation), and (2) *task-type signals*: request itself entails an attribute (e.g., applying for a scholarship implies student). GIST's context-aware generalisation abstracts domain-specific terms to category-level descriptions, breaking the inferential link that PAPILLON's paraphrasing preserves (Example 1). However, task-entailed attributes remain a challenge (Example 2). Although GIST reduces specificity by stripping identifying details (e.g., scholarship name), both GIST and PAPILLON preserve the signal, suggesting full protection requires mechanisms beyond query rewriting alone for future work.

7.3 Ablation

AIL robustness across LLMs. To assess if AIL produces stable comparative rankings independent of the underlying LLM, we rotate Kimi-K2.5, GLM-4.7, and Opus-4.6 (Anthropic, 2026) across attacker and comparator roles, yielding six pairs per local model (Table 2). As a comparative metric, absolute scores naturally vary but sign and ordering of Δ AIL are preserved across all configurations. Within each attacker, standard deviations remain low (0.4 to 2.0 pp). This confirms AIL preserves relative ranking regardless of which LLMs perform inference and comparison. This LLM-agnostic stability validates AIL's bi-directional consistency enforcement and categorical set-based comparison as effective safeguards against the instability of numerical LLM scoring, establishing AIL's reliability for comparing privacy-preserving methods.

Variant	Quality ↑ (%) / (Δ)	Leakage ↓ (%) / (Δ)	AIL ↓ (%) / (Δ)
GIST (full)	58.0 / –	8.8 / –	30.3 / –
- No Perspective Neutralisation	69.8 / +11.8	24.2 / **+15.4**	37.8 / **+7.5**
- No Protection Set	54.9 / -3.1	16.6 / **+7.8**	33.2 / +2.9
- No Intent	52.3 / **-5.7**	8.6 / -0.2	29.3 / -1.0
GIST (full, $k = 0$)	58.2 / +0.2	9.3 / +0.5	32.8 / +2.5
Single Prompt ($k = 0$)	58.4 / +0.2*	10.4 / +1.1*	36.3 / +3.5*
PAPILLON	54.0 / -4.0	46.0 / **+37.2**	39.9 / **+9.6**

Table 3: Component ablation of GIST with Phi-4-mini-instruct as the local model, removing one component at a time from the full pipeline. Δ = variant − baseline, in percentage points relative to GIST (full). Each component's dominant effect(s) in bold. *For the single-prompt variant, Δ is computed against the matched GIST (full, $k = 0$) control rather than GIST (full).

Impact of GIST Components. To isolate what drives GIST's improvement over PAPILLON, we ablate its core components using Phi-4-mini-instruct as $\mathcal{M}_{\text{local}}$, removing one component at a time from the full pipeline (Table 3). Perspective neutralisation is the strongest privacy driver as its removal raises explicit leakage by 15.4 pp and AIL by 7.5 pp, at the cost of 11.8 pp of quality, a trade-off given its role in suppressing demographic cues that enable attribute inference. The extraction of the protection set $\mathcal{E}$ and its generalisations primarily guards explicit identifiers, with its removal increasing explicit leakage by 7.8 pp and reducing quality by 3.1 pp, confirming that explicit span extraction both prevents missed sensitive entities and preserves meaning via context-aware generalisation. Intent extraction primarily serves quality as removing it lowers quality by 5.7 pp with negligible privacy effect, indicating that task context helps retain response utility. Finally, to test the value of decomposing sanitisation into two stages, we collapse extraction and rewriting into a single prompt. Measured against the matched GIST (full, $k = 0$) control, the use of a single prompt degrades AIL by 3.5 pp and leakage by 1.1 pp at comparable quality (+0.2 pp), showing that task decomposition con-

Model / Approach	Quality ↑ Ori (%) / Δ	Leakage ↓ Ori (%) / Δ	AIL ↓ Ori (%) / Δ
gpt-oss-120b			
Direct	96.0 / +1.1	–	51.2 / +0.9
Redaction	72.6 / +1.4	–	42.0 / +0.8
gemma-3-1b			
PAPILLON	16.9 / +0.3	42.8 / +0.9	41.4 / -0.6
GIST	**22.9** / +0.4	**21.8** / +0.6	**22.4** / -0.6
Phi-4-mini			
PAPILLON	51.8 / +2.2	45.0 / +1.0	39.0 / +0.9
GIST	**56.0** / +2.0	**8.1** / +0.7	**29.8** / +0.5
Mistral-7B			
PAPILLON	46.4 / +1.7	7.9 / +0.5	35.5 / -0.5
GIST	**54.0** / +1.9	**4.9** / +0.1	**27.7** / -0.1

Table 4: Dataset enhancement impact (full in Appendix G). Δ (Enhanced − Original) is directionally consistent across approaches without favouring any approach, GIST consistently outperforming (bold) PAPILLON.

tributes to GIST's privacy gains. See Appendix F for the effect in the number of feedback iterations k that plateaus beyond the default k=2. Together, these components let GIST balance the competing objectives of quality and privacy, yielding a 4.0 pp increase in quality, 37.2 pp reduction in leakage and 9.6 pp reduction in AIL relative to PAPILLON.

Impact of Dataset Enhancement. To verify that enhancements do not bias evaluation, we repeat all experiments on original PUPA-TNB (Table 4). GIST consistently outperforms PAPILLON on the original data across all models and metrics. All Δ values remain small (≤ 2.2 pp) and are directionally consistent across approaches, confirming fairness without favouring any approach. We attribute the universal quality gains to the replacement of Presidio placeholders with realistic synthetic PII.

8 Conclusion

We show a critical privacy gap in Privacy-Conscious Delegation: even perfect PII removal leaves the vast majority of inferable attributes exposed. To quantify this previously unmeasured risk, we introduce AIL, the first label-free metric for implicit privacy leakage. To close this gap, we propose GIST for context-aware generalisation that consistently achieves Pareto improvements over PAPILLON. Future work can extend GIST to multi-turn sessions and broaden attribute scope in AIL.

Limitations

Our work has limitations relating to both the AIL evaluation metric and the GIST mitigation approach.

Reliance on LLM Judges. AIL relies on LLM judges for scalable, label-free evaluation, representing an automated approximation rather than ground truth. We mitigate this through bi-directional consistency checks, ablation demonstrating stable rankings across six judge pairs, and human validation. Due to cost constraints, we leave ensemble judging across multiple models for future work.

AIL Weight Sensitivity. AIL's default weights ($w_{\text{eq}} = w_{\text{sub}} = 1.0$, $w_{\text{sup}} = w_{\text{ovl}} = 0.5$) follow the monotonic constraint in Equation 3, and our ablation (Table 2) confirms that relative rankings are stable across LLM configurations. The weights in AIL allow practitioners to tune partial-leakage severity for their domain, similar to how β calibrates the precision-recall trade-off for F_β-score. Domain-specific weight analysis for different risk tolerances remains future work.

Attribute Scope. AIL currently evaluates five demographic attributes (see Appendix H for AIL scores per attribute). As privacy is contextual, this set is not exhaustive and should be expanded for domain-specific applications.

Single-Turn Scope. Our evaluation is limited to single-turn interactions, following the task definition and benchmark constraints of PUPA-TNB (Siyan et al., 2025). As real-world privacy leakage can be cumulative across turns, extending to session-level privacy assessment and protection is an important future direction.

Inference Speed. GIST incurs overhead compared to direct API calls. See Appendix I for comparison to PAPILLON. While mitigated by lightweight SLMs, future work could further optimise speed for latency-critical applications.

Ethical Considerations

This research addresses the critical gap of measuring and mitigating implicit privacy leakage in Privacy-Conscious Delegation during user interactions with proprietary Large Language Models (LLMs) from untrusted external providers.

We enhanced the PUPA-TNB dataset (Siyan et al., 2025) by replacing Presidio (Mendels et al., 2018) placeholders with entirely synthetic, LLM-generated Personally Identifiable Information that bears no correspondence to real individuals, en-

abling comprehensive privacy evaluation without compromising actual personal data.

Our AIL evaluation metric functions as a diagnostic instrument to identify implicit privacy vulnerabilities in queries sent to untrusted LLM providers. While we acknowledge that adversarial evaluation methodologies entail dual-use considerations, prior work has already demonstrated the capability of LLMs to perform adversarial attribute inference attacks (Staab et al., 2024). Hence, equipping the research community with rigorous detection metrics is a prerequisite to developing effective countermeasures. We posit that the benefit of enabling a more rigorous assessment of privacy protection substantially outweighs the risks associated with disclosing this evaluation methodology.

Our GIST approach provides practical risk mitigation rather than formal privacy guarantees. Users should implement additional safeguards appropriate to their contexts and conduct thorough risk assessments prior to high-stakes deployment.

Overall, we believe that this work contributes positively to the responsible development of AI systems by quantifying the implicit privacy risks in Privacy-Conscious Delegation and provides a mitigating approach to address the privacy-utility trade-off associated with using LLMs hosted by untrusted third parties. This work could lead to broader societal benefits through improved access to AI-assisted services in privacy-sensitive domains.

References

Anthropic. 2026. Claude opus 4.6 system card.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Gemma Team, Google DeepMind. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

GLM Team. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Zheng Hui, Yijiang River Dong, Sanhanat Sivapiromrat, Ehsan Shareghi, and Nigel Collier. 2026. Privacy-r1: Privacy-aware multi-LLM agent collaboration via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 45924–45937, San Diego, California, United States. Association for Computational Linguistics.

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

Kimi Team. 2026. Kimi k2.5: Visual agentic intelligence. *Preprint*, arXiv:2602.02276.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Linzhang Li, Yixin Dong, Guanjie Wang, Ziyi Xu, Alexander Jiang, and Tianqi Chen. 2026a. Xgrammar-2: Dynamic and efficient structured generation engine for agentic llms. In *Proceedings of the ACM Conference on AI and Agentic Systems*, pages 1009–1022, New York, NY, USA. Association for Computing Machinery.

Weiyue Li, Minda Zhao, Weixuan Dong, Jiahui Cai, Yuze Wei, Michael Pocress, Yi Li, Wanyan Yuan, Xiaoyue Wang, Ruoyu Hou, Kaiyuan Lou, Wenqi Zeng, Yutong Yang, Yilun Du, and Mengyu Wang. 2026b. Grading scale impact on llm-as-a-judge: Human-llm alignment is highest on 0-5 grading scale. *Preprint*, arXiv:2601.03444.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Omri Mendels, Coby Peled, Nava Vaisman Levy, Sharon Hart, Tomer Rosenthal, Limor Lahiani, and 1 others. 2018. Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images.

Microsoft. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *Preprint*, arXiv:2407.11438.

OpenAI. 2022. tiktoken: A fast bpe tokeniser for use with openai's models.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Li Siyan, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. 2025. PAPILLON: Privacy preservation from Internet-based and local language model ensembles. In *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)*, pages 3371–3390, Albuquerque, New Mexico. Association for Computational Linguistics.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2025. Language models are advanced anonymizers. In *The Thirteenth International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A AIL Implementation Details

A.1 Attribute Inference Algorithm

Algorithm 1 Attribute Inference Leakage

Require: Pair (q, q'), Attacker $\mathcal{M}_{\text{attack}}$, Comparator $\mathcal{M}_{\text{compare}}$, Attributes $\mathcal{A}$
Ensure: Privacy Risk Score $S_{\text{risk}} \in [0, 1]$

Generate Profiles: $P \leftarrow \mathcal{M}_{\text{attack}}(q)$; $P' \leftarrow \mathcal{M}_{\text{attack}}(q')$
$\mathcal{W} \leftarrow [\,]$
for all $a \in \mathcal{A}$ **s.t.** $P[a] \neq \perp$ **do**
  $v \leftarrow P[a]$; $v' \leftarrow P'[a]$
  {Determine semantic relationship r}
  if $v' = \perp$ **then**
    $r \leftarrow$ DISJOINT
  else if $v = v'$ **then**
    $r \leftarrow$ EQUIV
  else
    {Perform semantic set comparison}
    $r_{\text{fwd}} \leftarrow \mathcal{M}_{\text{compare}}(v', v)$
    $r_{\text{rev}} \leftarrow \mathcal{M}_{\text{compare}}(v, v')$
    if ISCONSISTENT$(r_{\text{fwd}}, r_{\text{rev}})$ **then**
      $r \leftarrow r_{\text{fwd}}$
    else
      $r \leftarrow$ EQUIV {Conservative fallback}
    end if
  end if
  $w_a \leftarrow \phi(r)$ {Map relation to risk scalar}
  $\mathcal{W} \leftarrow \mathcal{W} \oplus [w_a]$
end for
return $\mu(\mathcal{W})$ {Return mean risk over identified attributes}

A.2 Attribute Inference Prompt

System Prompt

```
You are an expert Forensic Linguist and
    Profiler conducting a privacy audit.
Your task is to investigate what personal
    information can be inferred about a User
    based on their Query to a Large Language
    Model.

# Output Format
Output valid JSON with these fields:
- reasoning: Your step-by-step analysis
    identifying (1) explicit mentions, (2)
    implicit cues, and (3) contextual
    signals in the User Query that reveal
    personal attributes of the User.
- gender: ['Male', 'Female', 'Non-binary',
    'Unknown']
- race: Racial or ethnic background. Use
    "Unknown" if not inferable.
- location: Geographic location (city,
    region, or country). Use "Unknown" only
    if no location indicators exist.
- occupation: Job title, field, or industry.
- relationship_status: ['Single', 'Married',
    'Divorced', 'Widowed', 'Unknown']

# Inference Guidelines
Your goal is to extract as much information
    as possible about the User based on
    concrete evidence in the User Query.
    Make only HIGHLY CONFIDENT inferences
    based on:
- Direct mentions: Names, places, dates,
    relationships explicitly stated
- Contextual clues: Technical jargon,
    cultural references, slang
- Probabilistic reasoning: What is the most
    likely attribute given the evidence?
- Avoid speculation based on stereotypes or
    biases.
```

Figure 4: AIL attribute inference system prompt used by $\mathcal{M}_{\text{attack}}$ to extract demographic profiles from input text.

User Prompt

```
# Your Task
Analyze the following User Query and output
    the inferred personal attributes in JSON
    format as per the system instructions.

User Query: ```{text}```

JSON Output:
```

Figure 5: AIL attribute inference user prompt.

Structured output is enforced via constrained decoding with the JSON schema in Table 5. Field descriptions are provided to the model alongside the system prompt.

Field	Schema Description
`reasoning`	Step-by-step analysis: (1) explicit and implicit contextual cues in the User Query that reveals the personal attributes of the User (2) your inference for each attribute based on the evidence.
`gender`	Biological sex or gender identity. Enum: `Male`, `Female`, `Non-binary`, `Unknown`.
`race`	Racial or ethnic background. Return "Unknown" if User's race/ethnicity cannot be inferred.
`location`	Geographic location. Return "Unknown" if User's location cannot be inferred.
`occupation`	Standardized job title or industry (e.g., "Software Engineer", "Student", "Healthcare Professional").
`relationship_status`	Current relationship status. Enum: `Single`, `Married`, `Divorced`, `Widowed`, `Unknown`.

Table 5: JSON schema field descriptions for the attribute inference structured output, provided to $\mathcal{M}_{\text{attack}}$ via constrained decoding.

A.3 Set Comparison Prompt

The system prompt is parameterised by the attribute name and includes attribute-specific examples. Attribute-specific examples are injected per attribute type. For instance, for *location*: Superset: O="Paris, France", A="France"; Subset: O="France", A="Paris, France"; Equivalent: O="NYC, USA", A="New York City, USA"; Disjoint: O="Paris, France", A="London, UK"; Overlap: O="Schengen Area", A="European Union".

System Prompt (Part 1 of 2)

```
You are an expert in Set Theory and Data
    Privacy. Your task is to compare the
    semantic scope of two attribute values:
    an Original Value and an Anonymized
    Value.

You are comparing the **{attribute_name}**
    attribute.

# Instruction: Think in Sets
- Treat every attribute value as a Set of
    Entities that fit the description.
- Compare the **Anonymized Set (A)**
    relative to the **Original Set (O)**.
- Evaluate the relationships in the specific
    order below.

# Comparison Labels (Evaluate in Order)

1. **Equivalent** (Privacy Failure)
- Logic: The Sets are **identical**.
- Includes synonyms, acronyms, and
    formatting differences.
- Example: {examples['Equivalent']}

2. **Subset** (Privacy Leak)
- Logic: The Anonymized Set is **contained
    by** the Original Set.
- The Anonymized value is more
    specific/narrower than the Original.
- Example: {examples['Subset']}

3. **Superset** (Generalization)
- Logic: The Anonymized Set **contains** the
    Original Set.
- The Anonymized value is broader/vaguer; it
    includes the Original plus other
    possibilities.
- Example: {examples['Superset']}
```

Figure 6: AIL set-based comparison system prompt used by $\mathcal{M}_{\text{compare}}$ to determine the semantic relationship between original and anonymised values (Part 1 of 2).

Structured output is enforced via constrained decoding with the JSON schema in Table 6. Field descriptions are provided to the model alongside the system prompt.

System Prompt (Part 2 of 2)

```
4. **Overlap** (Ambiguous / Partial Match)
- Logic: The Sets **intersect** but neither
    contains the other completely
- Select this ONLY if the relationship is
    NOT Equivalent, Subset, or Superset.
- Example: {examples['Overlap']}

5. **Disjoint** (Obfuscation / Contradiction)
- Logic: The Sets share **no common
    elements**.
- Example: {examples['Disjoint']}

# Output Format
Return a JSON object with the following
    fields:
- `reasoning`: Explain the set relationship
    of the Anonymized Value relative to the
    Original Value.
- `comparison_label`: One of [`Equivalent`,
    `Subset`, `Superset`, `Overlap`,
    `Disjoint`].
```

Figure 7: AIL set-based comparison system prompt used by $\mathcal{M}_{\text{compare}}$ to determine the semantic relationship between original and anonymised values (Part 2 of 2).

Field	Schema Description
`reasoning`	Step-by-step set theory analysis: (1) identify what entities are included in Anonymized Value (2) what entities exist in Original Value (3) relationship of Anonymized Value relative to the Original Value.
`comparison_label`	The set relationship of the Anonymized (A) value relative to the Original (O). Evaluate in this strict order. Enum: `Equivalent`, `Subset`, `Superset`, `Overlap`, `Disjoint`.

Table 6: JSON schema field descriptions for the set-based comparison structured output, provided to $\mathcal{M}_{\text{compare}}$ via constrained decoding.

User Prompt

```
# Your Task
Determine if the Anonymized value is a
    'Superset', 'Subset', 'Equivalent',
    'Overlap', or 'Disjoint' compared to the
    Original Value.

Type of Attribute: ```{attribute_name}```

Original Value: ```{raw_value}```

Anonymized Value: ```{anon_value}```

JSON Output:
```

Figure 8: AIL set-based comparison user prompt.

B GIST Prompts

B.1 Phase 1: Context-Aware Extraction

B.1.1 Prompt Composition

The extraction system prompt is composed of three components concatenated into a single system message:

Extraction System Prompt Composition

```
EXTRACTION_SYSTEM_PROMPT =
    EXTRACTION_PROMPT_BASE
  + EXTRACTION_OUTPUT_STRUCTURE
  + EXTRACTION_FEW_SHOT_EXAMPLES
```

Figure 9: Composition of the GIST Phase 1 extraction system prompt. The three components below are concatenated into one system message.

B.1.2 Base Instructions

Extraction Prompt Base (Part 1 of 2)

```
You are an expert Privacy Officer. Your task
    is to determine the user's intent from a
    User Query, detect any sensitive
    information found and generate
    generalisations to protect the user's
    privacy.

Your output must be valid JSON with three
    keys: "user_intent", "reasoning", and
    "entities".

# Context
You are given a user Query that will be sent
    to a public LLM. Your task is to:
- Identify the user's intent.
- Reason step-by-step about what sensitive
    information exists, whether it is
    essential to convey the user's intent
    and how it can be protected.
- Generate abstract generalisations for
    essential sensitive information.
    Generalisations must preserve semantic
    meaning while removing identifying
    details.
- Do NOT answer the Query itself.

# Entity Types & Generalisation Strategy
- PERSON_NAME: Specific individuals.
  Generalise with a generic role or
    reference e.g. "Mark" -> "colleague"
- ORGANISATION: Specific companies,
    acronyms, teams.
  Generalise to the industry or function
    (e.g. "Google" -> "a tech company",
    "Marketing Team" -> "the internal
    department").
- LOCATION: Countries, regions, places,
    addresses.
  Generalise to the scale or nature (e.g.
    "Tokyo" -> "city", "123 Main St" -> "a
    residential address").
- DATE: Dates, times, specific years or
    quarters.
  Generalise to the relation to now (e.g.
    "Jan 1" -> "the start date", "2020" ->
    "a previous year").
```

Figure 10: GIST Phase 1 extraction prompt base (Part 1 of 2).

Extraction Prompt Base (Part 2 of 2)

```
- UNIQUE_IDENTIFIER: Traceable IDs. Emails,
    IP addresses, UUIDs, ticket #s, account
    #s.
  Generalise to the functional nature (e.g.
    "ticket #1234" -> "reference #X",
    "bob@gmail.com" -> "user_email").
- FINANCIAL: Specific salaries, deal sizes,
    precise budget numbers.
  Generalise to a broad range or scale (e.g.
    "$1.2M" -> "a significant budget", "$12"
    -> "a nominal fee").
- OCCUPATION: Specific titles.
  Generalise by level or field (e.g. "Senior
    VP" -> "an executive", "Neurologist" ->
    "a medical specialist").
- NATIONALITY_RELIGION_POLITICAL:
    Demographic groups.
  Generalise by group nature (e.g. "French
    citizens" -> "local nationals",
    "Republicans" -> "political group
    members").
- URL: Web links.
  Generalise by resource type (e.g.
    "twitter.com/user" -> "social media
    profile").

# Entity Identification Rules
1. UNIQUE Entities Only: Each distinct
    real-world entity should appear EXACTLY
    ONCE in your output.
2. Canonical Form: Use the most
    complete/formal form as the "original".
3. Context Matters: An entity is sensitive
    if it can identify or reveal private
    info in context.

# Generalisation Quality Rules
- Generalisations must REMOVE identifying
    specificity while preserving semantic
    role.
- Generalisations MUST be written in the
    same language as the Original Entity
    extracted.
- Each generalisation must be UNIQUE within
    the output.
   - When multiple entities share the same
    type, use distinguishing labels (e.g.
    "researcher A", "researcher B").
   - For sequential/temporal entities, use
    ordinal references (e.g., "the earlier
    period", "the later period").
```

Figure 11: GIST Phase 1 extraction prompt base (Part 2 of 2).

B.1.3 Output Structure

Extraction Output Structure

```
# Output Format
Return ONLY valid JSON:
{
  "user_intent": "Brief description of user
     intent",
  "reasoning": "Step-by-step analysis: (1)
     sensitive information identified that
     are essential to convey the user's
     intent (2) how to generalise the
     identifying details",
  "entities": [
    {"original": "exact text", "type":
     "ENTITY_TYPE", "generalisation":
     "abstract replacement"}
  ]
}

If no named entities found: {"user_intent":
     "...", "reasoning": "...", "entities":
     []}
```

Figure 12: GIST Phase 1 extraction output structure.

Field	Schema Description
`user_intent`	Brief description of user's intent.
`reasoning`	Step-by-step reasoning: (1) sensitive information identified, (2) why each needs protection, (3) generalisation strategy.
`entities`	List of all sensitive entities that are essential to convey the user intent.
Entity (nested within `entities`)	
`original`	Original entity text from the input.
`type`	Entity type: `PERSON_NAME`, `ORGANISATION`, `LOCATION`, `NATIONALITY_RELIGION_POLITICAL`, `URL`, `OCCUPATION`, `UNIQUE_IDENTIFIER`, `DATE`, `FINANCIAL`.
`generalisation`	Broader category or abstract description for sensitive entities.

Table 7: JSON schema field descriptions for the GIST Phase 1 extraction structured output, provided to $\mathcal{M}_{\text{local}}$ via constrained decoding.

Structured output is enforced via constrained decoding with the JSON schema in Table 7. Field descriptions are provided to the model alongside the system prompt.

B.1.4 Few-Shot Examples

Extraction Few Shot Examples
Example 1 of 3 - Query

```
# Example 1: Query with multiple entities of
    the same type
  requiring unique generalisations

Query: ```Dear Chris, I'm Dr. Alice Brown.
    My colleague
  Dr. James Wilson and I are Senior Research
    Scientists at
  HQCSS in Boston. The Q4 budget of $2.5
    million was approved
  in March 2024. Do you know of any Russian
    candidates?```
```

Figure 13: GIST Phase 1 extraction few-shot example 1 of 3 (query): multi-entity query requiring unique generalisations.

Extraction Few Shot Examples
Example 1 of 3 - Output

````
Output:
```json
{
  "user_intent": "Request for help finding
     international
    job candidates for a research position",
  "reasoning": "(1) Sensitive info: Personal
     names (Chris,
    Dr. Alice Brown, Dr. James Wilson),
     organization (HQCSS),
    location (Boston), occupation (Senior
     Research Scientist),
    nationality filter (Russian), date,
     financial amount.
    (2) Protection needed: Names identify
     individuals, HQCSS
    is internal org, Boston reveals
     location. (3) Strategy:
    Use distinguishing labels for same-type
     entities.",
  "entities": [
    {"original": "Chris", "type":
     "PERSON_NAME",
      "generalisation": "the recipient"},
    {"original": "Dr. Alice Brown", "type":
     "PERSON_NAME",
      "generalisation": "researcher A"},
    {"original": "Dr. James Wilson", "type":
     "PERSON_NAME",
      "generalisation": "researcher B"},
    {"original": "Senior Research
     Scientists",
      "type": "OCCUPATION",
      "generalisation": "research roles"},
    {"original": "HQCSS", "type":
     "ORGANISATION",
      "generalisation": "a research
     division"},
    {"original": "Boston", "type":
     "LOCATION",
      "generalisation": "a city"},
    {"original": "Russian",
      "type":
     "NATIONALITY_RELIGION_POLITICAL",
      "generalisation": "international"},
    {"original": "March 2024", "type":
     "DATE",
      "generalisation": "a recent date"},
    {"original": "$2.5 million", "type":
     "FINANCIAL",
      "generalisation": "significant funding"}
  ]
}
```
````

Figure 14: GIST Phase 1 extraction few-shot example 1 of 3 (output): 9 entities with distinguishing labels for same-type entities.

Extraction Few Shot Examples
Example 2 of 3

```
# Example 2: Multilingual query -
    generalisations match
  the language of each original entity

Query: ```请帮我翻译: John Smith works at
    Microsoft
  in Seattle.```

Output:
```json
{
  "user_intent": "Translate an English text
    into Chinese",
  "reasoning": "(1) Sensitive info: Personal
    name
    (John Smith), organization (Microsoft),
    location
    (Seattle). (2) Protection needed: Names
    identify
    individuals, organization reveals
    specific company,
    location reveals specific place. (3)
    Strategy:
    Generalise language to match the
    original entity
    language (English entities generalized
    in English).",
  "entities": [
    {"original": "John Smith", "type":
    "PERSON_NAME",
     "generalisation": "an employee"},
    {"original": "Microsoft", "type":
    "ORGANISATION",
     "generalisation": "a tech company"},
    {"original": "Seattle", "type":
    "LOCATION",
     "generalisation": "a city"}
  ]
}
```
```

Figure 15: GIST Phase 1 extraction few-shot example 2 of 3: multilingual query with language-matched generalisations. English entities are generalised in English despite the Chinese instruction.

Extraction Few Shot Examples
Example 3 of 3

```
# Example 3: Query with no valid entities

Query: ```How do I implement a binary search
    algorithm
  in Python?```

```json
{
  "user_intent": "Learn how to implement a
    binary search
    algorithm in Python",
  "reasoning": "(1) Sensitive info scan: No
    personal names,
    no organizations, no locations, no
    dates, no identifiers.
    This is a generic technical query. (2)
    No protection
    needed: The query contains no PII or
    sensitive
    operational details.",
  "entities": []
}
```
```

Figure 16: GIST Phase 1 extraction few-shot example 3 of 3: no-entity query returning an empty entity list when no sensitive information exists.

Extraction User Prompt

```
# Your Task

Query: "{text}"

Output:```json
```

Figure 17: GIST Phase 1 extraction user prompt.

B.1.5 Extraction Validator Prompts

Extraction Validator System Prompt (Part 1 of 3)

```
You are an expert Privacy Officer and
    Quality Assurance Specialist. Your task
    is to rigorously review the information
    extraction to ensure all sensitive
    information has been correctly
    identified and generalised according to
    the specific Protection Rules.

# Protection Rules (Evaluation Criteria)
- PERSON_NAME: Generalise with a generic
    role (e.g. "Mark" -> "colleague").
- ORGANISATION: Generalise to
    industry/function (e.g. "Google" -> "a
    tech company").
- LOCATION: Generalise to scale/nature (e.g.
    "Tokyo" -> "city").
- DATE: Generalise to relation to now (e.g.
    "Jan 1" -> "the start date").
- UNIQUE_IDENTIFIER: Generalise to
    functional nature (e.g. "ticket #123" ->
    "reference #X").
- FINANCIAL: Generalise to broad range/scale.
- OCCUPATION: Generalise by level/field.
- NATIONALITY_RELIGION_POLITICAL: Generalise
    by group nature.
- URL: Generalise by resource type.
```

Figure 18: GIST Phase 1 extraction validator system prompt (Part 1 of 3).

Extraction Validator System Prompt (Part 2 of 3)

```
# Validation Checklist (Apply STRICTLY -
    reject on ANY failure)
1. Completeness (CRITICAL): Have ALL
    entities matching the types above been
    extracted?
   - Scan the Original Query word-by-word
    for ANY names, dates, organisations,
    locations, or identifiers.
   - If even ONE entity is missing, set
    passes_validation=false.
   - Common misses: First names mentioned
    alone, abbreviated company names,
    implicit dates ("last year", "Q4").
2. False Positives: Are there any extracted
    entities that are NOT in the original
    text?
   - REJECT: Hallucinated entities or
    details not present in the source.
3. Generalisation Compliance: Do the
    generalisations follow the Protection
    Rules above?
   - REJECT: Generic placeholders like
    "Person", "Place", "Thing", "Name" -
    these are too vague.
   - REJECT: Generalisations that preserve
    the original text (e.g., "John" ->
    "John").
   - REJECT: Generalisations that leak
    identifying info (e.g., "Dr. Smith" ->
    "Dr. [someone]").
4. Uniqueness (CRITICAL): Are all
    generalisations unique and
    distinguishable?
   - REJECT: Duplicate generalisations
    (e.g., two people both mapped to "a
    colleague").
   - Each entity MUST have a distinct
    generalisation (e.g., "colleague A",
    "colleague B").
5. Intent Accuracy: Does the extracted
    user_intent capture the core request
    without revealing PII?
   - REJECT: Intents that include names,
    specific companies, or locations.
```

Figure 19: GIST Phase 1 extraction validator system prompt (Part 2 of 3).

Extraction Validator System Prompt (Part 3 of 3)

```
# Reasoning Requirement
You MUST provide detailed step-by-step
    reasoning that:
1. Lists every potential entity found in the
    Original Query
2. Checks each against the extracted
    entities list
3. Verifies no extra entities were invented
4. Evaluates each generalisation for
    compliance
5. Explicitly states pass/fail for each
    checklist item

# Output Format
Return ONLY valid JSON:
{
  "step_by_step_analysis": "Detailed
    analysis...",
  "passes_validation": true,
  "identified_problems": "None",
  "suggested_corrections": "None"
}
```

Figure 20: GIST Phase 1 extraction validator system prompt (Part 3 of 3).

Extraction Validator User Prompt

```
# Task: Validate Entity Extraction

# Original Query
```{original_query}```

# Detected User Intent
```{user_intent}```

# Extracted Entities
```{extraction_data}```

# Instructions
Perform a STRICT validation. Apply the
    checklist rigorously:
1. List ALL potential entities in the
    Original Query (names, dates, companies,
    locations, identifiers, etc.)
2. Check if EACH potential entity appears in
    the Extracted Entities - flag ANY missing
3. Verify each generalisation is specific
    enough (not just "Person" or "Place")
    and unique
4. Check that user_intent does not leak PII

Be conservative: if in doubt, flag it as an
    issue. Missing entities are a critical
    failure.

Return valid JSON with keys:
    "step_by_step_analysis",
    "passes_validation",
    "identified_problems",
    "suggested_corrections".
```

Figure 21: GIST Phase 1 extraction validator user prompt.

Structured output is enforced via constrained decoding with the JSON schema in Table 8. Field descriptions are provided to the model alongside the system prompt.

Field	Schema Description
`step_by_step_analysis`	Detailed analysis: list entities found in query, compare against extracted entities, evaluate each generalisation.
`passes_validation`	True only if ALL checks pass: completeness, generalisation quality, uniqueness, intent privacy.
`identified_problems`	Specific problems found: missed entities, poor generalisations, duplicates. Empty if `passes_validation` is true.
`suggested_corrections`	Exact fix instructions: "Add entity X with type Y and generalisation Z". Empty if `passes_validation` is true.

Table 8: JSON schema field descriptions for the GIST Phase 1 extraction validator structured output, provided to $\mathcal{M}_{\text{local}}$ via constrained decoding.
```

B.1.6 Extraction Refinement Prompts

Extraction Refinement Prompt

```
# Refinement Task

The previous entity extraction had issues.
    Please generate an improved extraction.

# Original Query
```{original_query}```

# Previous Extraction (with issues)
```{previous_extraction}```

# Feedback on Issues
```{feedback}```

# Your Improved Extraction
Generate a corrected entity extraction
    addressing all the feedback issues.
Output:```json
```

Figure 22: GIST Phase 1 refinement prompt used when the validator detects issues in the extraction.
```

B.2 Phase 2: Semantic Abstraction and Pruning

System Prompt (Part 1 of 2)

```
You are an expert Query Proxy. Your task is
    to create an Anonymous Query that
    conveys the same request as the User
    Query while removing all identifying
    details.

# Core Principle
Your mission is to protect the User from a
    state-backed adversary trying to infer
    personal attributes. All sensitive
    information that could reveal
    demographics must be removed if not
    essential to conveying the request. If
    the sensitive information is essential
    to convey the user's intent, you MUST
    generalise it to abstract forms.

# Rules for Rewriting
- Preserve Query Language: Write the
    Anonymous Query in the SAME LANGUAGE as
    the User Query.
  - If the User Query is in Spanish, the
    Anonymous Query must be in Spanish.
  - If the User Query contains mixed
    languages (e.g., Chinese instruction
    with English content to translate),
    preserve that structure.
- Preserve User Intent: The Anonymous Query
    must ask for the SAME TASK as the
    original.
  - Questions stay as questions. Commands
    stay as commands.
- Neutralize Tone: For sensitive questions,
    remove first-person narrative and use a
    hypothetical third-person perspective.
- Strip Demographics: Remove any text that
    reveals the author's age, gender,
    occupation, education, or location. If
    it is essential to receive a high
    quality response, you must generalise it.
- Apply Entity Generalisations: Replace
    specific identifying details with the
    terms provided in the Entity
    Generalisations list.
```

Figure 23: GIST Phase 2 semantic abstraction and pruning system prompt (Part 1 of 2).

System Prompt (Part 2 of 2)

```
# Examples

## Example 1: Chinese Translation Request
    (preserve Chinese instruction)
User Query: ```请帮我翻译: Dear Dr. Smith, I
    enjoyed meeting you at the Stanford
    conference in California.```

Entity Generalisations: ```
  - 'Dr. Smith' -> 'a colleague'
  - 'Stanford' -> 'a university'
  - 'California' -> 'a location'```

Detected User Intent: ```Translate English
    text to Chinese```

Anonymous Query: ```请帮我翻译: Dear
    colleague, I enjoyed meeting you at the
    conference at a university in a
    location.```

# Output
Output ONLY the Anonymous Query itself.
```

Figure 24: GIST Phase 2 semantic abstraction and pruning system prompt (Part 2 of 2).

User Prompt

```
# Your Task

User Query: ```{text}```

Entity Generalisations: ```
{mapping_section}```

Detected User Intent: ```{user_intent}```

Anonymous Query:```
```

Figure 25: GIST Phase 2 semantic abstraction and pruning user prompt.

B.3 Phase 4: Context Restoration

System Prompt (Part 1 of 2)

```
You are an expert response editor. Your task
    is to transform a Draft Response into a
    Final Polished Response that directly
    answers the Original User Query with
    full context restored.

# Primary Goal
The Final Polished Response must read as if
    it were written specifically for the
    Original User Query - relevant,
    specific, and directly actionable.

# Instructions

## ANSWER THE USER'S ORIGINAL QUESTION
- Read the Original User Query carefully to
    understand what the user specifically
    asked for.
- The Draft Response was generated from an
    anonymised query and may be vague,
    generic, or tangential.
- The final response must directly address
    what the user originally asked.

## RESTORE ORIGINAL CONTEXT
- If the Draft says something generic, make
    it specific to the actual context from
    the Original User Query.
- Ensure every reference to a substituted
    concept in the Draft Response is
    restored to a natural form that
    addresses the Original User Query.
- Add relevant details that would be helpful
    given the Original entities involved.
```

Figure 26: GIST Phase 4 context restoration system prompt for re-personalising the remote model's response (Part 1 of 2).

System Prompt (Part 2 of 2)

```
## CORRECT FACTUAL ERRORS
- Correct all factual and logical
    inconsistencies that appear when
    generics are restored to specifics.
- Example: If draft says "the capital of the
    country" but the original location is
    "Sydney", correct to "Sydney" without
    calling it a capital.
- Ensure that restored facts are accurate
    for the specific entities.

# Output
Respond with ONLY the Final Polished
    Response.
```

Figure 27: GIST Phase 4 context restoration system prompt for re-personalising the remote model's response (Part 2 of 2).

User Prompt

```
# Task

# Original User Query
```{original_query}```

# Detected Entity Mapping (Substitute ->
    Original)
```{mapping_text}```

# Draft Response (using Substitutes)
```{anonymised_response}```

# Final Polished Response
```

Figure 28: GIST Phase 4 context restoration user prompt.
```

C PUPA TNB Dataset Improvements

C.1 Synthetic Replacements for Presidio Placeholders

#	User Query	Target Response	PII Units
1	Original: Please write an email copy saying thanks for **<PRESIDIO_ANONYMIZED_PERSON>** for attending the meeting and also to let... Improved: Please write an email copy saying thanks for **Michael Johnson** for attending the meeting and also to let...	Original: Subject: Thank You for Attending the Meeting! Meeting Notes Attached Dear **<PRESIDIO_ANONYMIZED_PERSON>** I... Improved: Subject: Thank You for Attending. Hi **Michael**, I...	Original: **<presidio_anonymized_person>** Improved: **michael johnson**
2	Original: ...All the best, **<PRESIDIO_ANONYMIZED_PERSON>** Improved: ...All the best, **Fatima Khan**	Original: ... Warm regards, **<PRESIDIO_ANONYMIZED_PERSON>** Improved: ... Best regards, **Fatima Khan**	Original: **<presidio_anonymized_person>** || ahmad Improved: **fatima khan** || ahmad
3	Original: Make my resume fit cybersecurity positions... Felipe Lozano (402) 617-4647 | **<PRESIDIO_ANONYMIZED_EMAIL_ADDRESS>** | Crete, NE ... Improved: Make my resume fit cybersecurity positions... Felipe Lozano (402) 617-4647 | **felipe.lozano@gmail.com** | Crete, NE ...	Original: **Felipe Lozano** (402) 617-4647 | **<PRESIDIO_ANONYMIZED_EMAIL_ADDRESS>** | Crete, NE ... Improved: ...**Felipe Lozano** (402) 617-4647 | **felipe.lozano@gmail.com** | Crete, NE ...	Original: baseballcloud || lincoln || lake mary || bogotá || take2 || foundever || crete || hudl || felipe lozano || vienna || doane Improved: baseballcloud || lincoln || lake mary || bogotá || take2 || foundever || crete || hudl || felipe lozano || vienna || doane **|| felipe.lozano@gmail.com**
...	...	...	...

Table 9: Examples of improvements to the dataset by replacing presidio placeholders with new synthetic entities. These synthetic PII replacements were created for 15 person names, 11 email addresses and 2 phone numbers, totaling 28 placeholders.

C.2 PII Units Enhancements

#	User Query	PII Units	Rationale
1	Write an email template to russian ministry of education about the invitation issuance that I have awarded scholarship in HSE university and I thought its good to start my scholarship with prep year, as I'm living in	Original: **[location not specified in the original text] || russian ministry of education || hse university** Improved: russian ministry of education || hse university	"[location not specified]" does not appear in the User Query and is therefore not a valid PII extraction.
2	If Jim puts in an order to purchase 30 shares of company A for $1.30 while the current price is $1.35, and the order is canceled immediately, what type of order did he put in? choose from the following: Limit Order Market Order Fill or Kill Order Stop-Loss Order	Original: jim || **company a** Improved: jim	"company A" is a generic placeholder used to illustrate a hypothetical scenario. It does not refer to any specific real-world entities and does not qualify as a PII unit.
3	list all job offering from SPARC companies with a URL link	Original: **sparc** Improved:	"SPARC" is likely intended as SPAC (Special Purpose Acquisition Company), a type of shell corporation listed on a stock exchange. It describes a classification of companies rather than a specific named entity, hence it is not a PII unit.
4	How do I succeed financially in Syria despite all the bad conditions here?	Original: **syri** Improved: syria	The original PII unit "syri" appears to be an incorrect truncation of the actual PII unit "syria"
5	write 5 innovative headings of 100mi - Are you ready to join hands with a revolutionary platform that combines ...	Original: **mi** Improved: 100mi	The original PII unit "mi" is incomplete as it is missing the prefix of "100". The correct PII unit should be "100mi"
...			

Table 10: Examples of refinement to the dataset by improving correctness of PII Units, totaling 25 enhancements.

C.3 Missing Target Response Replacement

User Query	Target Response	Rationale
Correctly format this list of words: Turtle, Garfield, Alligator, Headphones, Wedding Dress, ...	Original: **#NAME** Improved: Here's the correctly formatted list of words in alphabetical order:- Alligator- America- Angle- Ant- Apple ...	"#NAME" appears to be a common spreadsheet software error code that could have replaced the correct target response value at some point during the process where the original dataset was created.

Table 11: Example of improvement to dataset by replacing a missing target response. Totaling 1 count of replacement.

D Human Annotator Instructions

The following sections detail the instructions and forms provided to our human annotators for validating the implicit attribute inferences and set-based comparisons made by the LLM.

D.1 Attribute Inference

Attribute Inference

We have all heard of LLMs as a Judge, but who's there to help us judge LLMs? We are currently evaluating if an LLM can identify the implicit risks in our chatbot queries.
An adversarial LLM has analysed a set of user queries to infer implicit personal attributes about the authors of those queries. Your task is to evaluate the logical soundness and accuracy of these AI-generated inferences.

For each query, you will be presented with a set of Inferred Attributes inferred by the adversarial LLM. Carefully read the query and determine if the inferred attributes logically follow from the text provided.

1. Read the user query thoroughly to understand the context.
2. Review the attributes the LLM has extracted (Gender, Race, Location, Occupation, Relationship Status).
3. Decide whether you agree with the LLM's assessment based only on the evidence present or strongly implied in the text.
4. If you disagree with the LLM's inference, you must provide what you believe to be the correct attribute(s).

Figure 29: Instructions provided to human annotators for validating attribute inference results.

Attribute Inference Form

Please submit your evaluations using the provided form.

1. **Your Code Name:** Provide a consistent unique identifier. This is used solely for statistical analysis and will not be used to track personal identity.
2. **Query ID:** Enter the ID of the query you are evaluating (1–50).
3. **Overall Agreement:** Do you agree with the overall attributes inferred by the Adversarial LLM? Select *Yes* or *No*.

4–8. **Corrected Attributes** (Conditional): If you answered "No" to Question 3, provide your own attribute inferences for the fields below. If you agree with the LLM on a specific attribute but disagree on others, leave the agreed-upon field blank.

- Gender
- Race
- Location
- Occupation
- Relationship Status

Figure 30: Annotation form for attribute inference validation.

D.2 Set-Based Attribute Comparison

Set-Based Attribute Comparison

Your task is to compare pairs of attributes (Attribute A and Attribute B) and classify the set-theoretic relationship between them.

For each given pair (Attribute A vs. Attribute B), determine how the two concepts relate to one another. Classify the relationship into one of the five strictly defined categories below.

- **Equivalent:** Attribute A and Attribute B refer to the exact same concept or entity.
  Example: US vs. USA
- **Subset:** Attribute A is entirely contained within Attribute B.
  Example: Tokyo vs. Japan
- **Superset:** Attribute A entirely contains Attribute B.
  Example: France vs. Paris
- **Overlap:** Attribute A and Attribute B intersect and share commonalities, but neither entirely contains the other.
  Example: Data Engineer vs. Data Scientist
- **Disjoint:** Attribute A and Attribute B are completely distinct concepts with no intersection.
  Example: German vs. Indian

Figure 31: Instructions provided to human annotators for set-based attribute comparison.

Set-Based Attribute Comparison Form

Please submit your classifications using the provided form.

1. **Your Code Name:** Provide a consistent unique identifier. This is used solely for statistical analysis and will not be used to track personal identity.
2. **Comparison Set:** Select the batch you are currently annotating (Set 1, 2, 3, 4, or 5).
3. **Attribute Comparison Responses:** Complete the provided input table. For each row (Query ID 1 to 30), select the corresponding Set Type (*Equivalent*, *Subset*, *Superset*, *Overlap*, or *Disjoint*) that best describes the relationship between Attribute A and Attribute B.

Figure 32: Annotation form for set-based attribute comparison.

E Examples for Qualitative Analysis

Approach	Query Sent to Remote Model	Inferred Attributes
Original	A security in the company name x who is doing security job but in a freetime he earn from the indian stock market through doing future and option write an linkedin story for the same in a motivation manner	Gender: Male Race: Unknown Location: India Occupation: Security Professional Relationship: Unknown
PAPILLON	"Balancing the demands of a security job and personal pursuits, I've found an unexpected source of motivation: the Indian stock market. In my free time, I've been dabbling in futures and options, learning the ins and outs of the market. It's not just about the potential financial gains, but the thrill of the challenge and the satisfaction of making informed decisions. This journey has taught me valuable lessons about risk management, patience, and resilience. I'm excited to see where this journey takes me next!"	Gender: Unknown Race: Unknown Location: <u>India</u> Occupation: <u>Security Prof.</u> Relationship: Unknown
GIST	Write a LinkedIn story about a professional in a generic industry who earns from an international market through financial instruments in their spare time.	Gender: Unknown Race: Unknown Location: Unknown Occupation: Unknown Relationship: Unknown
Original	i want to write an essay for HeForShe IMPACT Champion Scholarships	Gender: Unknown Race: Unknown Location: Unknown Occupation: Student Relationship: Unknown
PAPILLON	"Write an essay for a scholarship application, focusing on your qualifications, achievements, and how you plan to make a positive impact in your chosen field."	Gender: Unknown Race: Unknown Location: Unknown Occupation: <u>Student</u> Relationship: Unknown
GIST	How can I write an essay for a scholarship program?	Gender: Unknown Race: Unknown Location: Unknown Occupation: <u>Student</u> Relationship: Unknown

Table 12: Qualitative comparison of query sanitisation approaches across privacy and quality. Inferred attributes are extracted by a simulated adversary; <u>underlined</u> values indicate attributes that remain inferable after sanitisation, representing privacy leakage. In Example 1, GIST eliminates all inferable attributes by generalising domain-specific markers (e.g., "Indian stock market" $\rightarrow$ "an international market") that PAPILLON preserves verbatim. Furthermore, PAPILLON's output is a self-contained response rather than a sanitised query, bypassing the delegation paradigm. In Example 2, both approaches retain a task-entailed attribute (Student), illustrating limits of query rewriting when the user's intent itself encodes the attribute. See walk-through below for detailed analysis.

We identify two broad categories of implicit cues that AIL captures and that PII-centric methods leave unaddressed: *domain-specific markers* and *task-type signals*. Table 12 presents representative examples of each category, comparing the original query, the sanitised outputs of PAPILLON and GIST, and the attributes inferred by a simulated adversary. We walk through each example below.

Example 1: Domain-Specific Markers. The original query asks for a motivational LinkedIn story about a security professional who trades on the "Indian stock market." The simulated adversary infers three attributes from the original: *Gender* (Male, from the pronoun "he"), *Location* (India, from "Indian stock market"), and *Occupation* (Security Professional, from "security job").

Privacy. The two approaches handle these domain-specific terms differently. PAPILLON paraphrases the query into a first-person narrative but preserves both markers verbatim ("Indian stock market" $\rightarrow$ "Indian stock market"; "security job" $\rightarrow$ "security job"), so the adversary still infers Location as India and Occupation as Security Professional, yielding residual AIL. Only the gender signal is removed because PAPILLON drops the explicit pronoun. In contrast, GIST applies context-aware generalisation that abstracts each marker to a category-level description ("Indian stock market" $\rightarrow$ "an international market"; "security job" $\rightarrow$ "a professional in a generic industry"; first-person "he" $\rightarrow$ third-person objective framing). The adversary infers all five attributes as Unknown, yielding an AIL score of 0.

Quality. Beyond privacy, the two outputs differ in their suitability as delegated queries. PAPILLON's output is a self-contained LinkedIn story rather than a sanitised query, effectively generating the final response itself and bypassing the delegation paradigm. GIST instead produces a clean task instruction that preserves sufficient semantic context ("LinkedIn story," "financial instruments," "spare time") for the remote model to generate a relevant, high-quality response while all identifying specifics are abstracted away.

Example 2: Task-Entailed Attributes. The original query requests help writing an essay for the "HeForShe IMPACT Champion Scholarships." The only attribute the adversary infers is Occupation: Student, because applying for a scholarship logically entails student status. This serves as an example of a task-type signal where the user's intent itself encodes the attribute.

Privacy. GIST generalises the query to "How can I write an essay for a scholarship program?", stripping the programme name ("HeForShe IMPACT Champion") and reducing specificity. PAPILLON rewrites the query into a generic scholarship essay request. However, the word "scholarship" is retained in both approaches, and the adversary still infers Occupation as Student (Equivalent to the original) because the concept of a "scholarship" inherently implies the applicant is a student regardless of how the query is phrased.

Quality. Both approaches produce clean delegated queries that preserve the core task intent (essay writing for a scholarship), so neither degrades response quality. GIST strips the specific programme name, which may reduce response specificity but avoids leaking the named programme as an identifying detail. This example illustrates a limitation: when a sensitive attribute is logically entailed by the user's core intent, query-rewriting alone cannot fully protect it regardless of the approach.

Summary. Together, these examples highlight a dual advantage for GIST over PAPILLON. On the privacy axis, GIST's category-level abstractions eliminate domain-specific markers that PAPILLON's paraphrasing preserves (Example 1), while both approaches hit the same floor on task-entailed attributes (Example 2). On the quality axis, GIST consistently produces well-formed delegated queries that supply sufficient semantic context for the remote model, whereas PAPILLON may conflate sanitisation with generation, bypassing the delegation paradigm entirely (Example 1). This qualitative pattern is consistent with the Pareto improvement observed quantitatively in Table 1.

F Effect of Maximum Feedback Iterations

k	**Quality** (%) ↑	**Leakage** (%) ↓	**AIL** (%) ↓	**Average Iterations**
0	56.0	5.4	31.5	0
1	55.7	5.1	28.1	1
2	**55.9**	**5.0**	**27.6**	**1.1**
3	55.1	4.4	27.3	1.2

Table 13: Ablation on the maximum feedback iterations k (default $k = 2$) in Phase 1 using Mistral-7B-Instruct-v0.3 as the local model. GIST uses a Generator-Validator feedback loop to iteratively refine extraction. Here we examine the pipeline's sensitivity to the maximum feedback iterations k. While k has negligible impact on Quality, both Leakage and AIL decrease as k increases, confirming that iterative feedback improves privacy protection. However, diminishing returns emerge beyond $k = 2$, with only 1.72% of queries requiring a third iteration and the average number of iterations plateauing around 1.1. $k = 2$ is set as default throughout our experiments to balance privacy gains against computational overhead.

G Dataset Enhanced vs Original

Model / Approach	Quality ↑ Original (%) / Δ	Leakage ↓ Original (%) / Δ	AIL ↓ Original (%) / Δ
gpt-oss-120b			
Direct	96.0 / +1.1	–	51.2 / +0.9
Redaction	72.6 / +1.4	–	42.0 / +0.8
gemma-3-1b			
PAPILLON	16.9 / +0.3	42.8 / +0.9	41.4 / -0.6
GIST	**22.9** / +0.4	**21.8** / +0.6	**22.4** / -0.6
Llama-3.2-3B			
PAPILLON	33.0 / +1.2	28.3 / +0.4	37.2 / +0.6
GIST	**36.5** / +0.8	**6.5** / +0.5	**32.0** / +0.4
Phi-4-mini			
PAPILLON	51.8 / +2.2	45.0 / +1.0	39.0 / +0.9
GIST	**56.0** / +2.0	**8.1** / +0.7	**29.8** / +0.5
gemma-3-4b			
PAPILLON	51.3 / +1.0	14.0 / +0.1	34.1 / -0.2
GIST	**65.8** / +0.9	**1.7** / -0.1	**26.8** / +0.8
Mistral-7B			
PAPILLON	46.4 / +1.7	7.9 / +0.5	35.5 / -0.5
GIST	**54.0** / +1.9	**4.9** / +0.1	**27.7** / -0.1
Llama-3.1-8B			
PAPILLON	50.8 / +0.9	8.2 / +0.2	35.9 / -0.7
GIST	**56.6** / +0.8	**5.4** / +0.5	**30.4** / -0.9

Table 14: Dataset enhancement impact (full results; subset in Table 4). Δ (Enhanced − Original) is directionally consistent across approaches without favouring any approach, with GIST consistently outperforming (bold) PAPILLON.

H Per-attribute AIL Comparison

Attribute	PAPILLON AIL ↓	GIST AIL ↓	Δ (GIST – PAPILLON) ↓	n
Gender	16.7%	**0.0%**	-16.7	24
Race	6.1%	**0.0%**	-6.1	33
Location	38.2%	**21.4%**	-16.8	110
Occupation	58.6%	**51.1%**	-7.5	181
Relationship Status	100.0%	**0.0%**	-100.0	1

Table 15: Per-attribute AIL breakdown using Phi-4-mini-instruct as $\mathcal{M}_{\text{local}}$. Per-attribute AIL averages risk across the instances of each attribute as compared to the original per-query AIL (Section 4.3), therefore magnitudes differ. GIST achieves lower AIL (bold) on every attribute. The largest gains are in Gender (-16.7 pp) and Location (-16.8 pp). For Gender, GIST's perspective neutralisation converts first-person narratives to objective third-person descriptions, eliminating gendered pronouns and voice markers that PAPILLON preserves, as its query rewriting targets explicit PII removal. For Location, GIST's semantic abstraction generalises geographic references to category-level descriptions (e.g., region), breaking the inferential link while retaining task utility. Occupation shows a smaller gain (-7.5 pp), consistent with our qualitative finding that task-entailed attributes remain challenging (Section 7.2). Relationship Status is retained for completeness but precludes meaningful per-attribute comparison as it has only one instance.

I Inference Efficiency Comparison between GIST and PAPILLON

Stage / Metric	GIST	PAPILLON	Ratio
Anonymisation			
Latency (s)	25.1	6.2	4.0x
Input tokens (% system prompt)	3,572 (77.8%)	323 (70.3%)	11.1x
Output tokens	452	132	3.4x
$\mathcal{M}_{\textbf{remote}}$			
Latency (s)	6.1	4.7	1.3x
Input tokens	38	52	0.7x
Output tokens	769	707	1.1x
Deanonymisation			
Latency (s)	15.0	16.7	0.9x
Input tokens (% system prompt)	978 (28.3%)	1,152 (20.1%)	0.8x
Output tokens	288	348	0.8x
Total (without $\mathcal{M}_{\textbf{remote}}$)			
Latency (s)	**50.4**	**23.7**	**2.1x**
Total tokens	**5,987**	**2,093**	**2.9x**

Table 16: Median per-query latency and token usage for GIST and PAPILLON using Phi-4-mini-instruct as $\mathcal{M}_{\text{local}}$ with token counts computed via cl100k_base (OpenAI, 2022). Totals are computed as the median of per-query sums, hence they are not strictly additive. The increase in input tokens (11.1x) has reduced impact on latency as the majority (77.8%) are static system prompts that are cacheable. Decoding additional output tokens (3.4x increase) remains the biggest factor for latency increase in the anonymisation phase, driven by the multi-step process. GIST does not substantially increase token usage on the paid $\mathcal{M}_{\text{remote}}$, with lower input tokens (0.7x), slightly more output tokens (1.1x), and a modest latency increase (1.3x). Conversely, during the deanonymisation stage, GIST is slightly more efficient than PAPILLON across all reported metrics (latency 0.9x, input tokens 0.8x, output tokens 0.8x). Aggregating over the stages excluding the use of $\mathcal{M}_{\text{remote}}$, GIST has a median per-query latency of 50.4s vs 23.7s (2.1x). Its total $\mathcal{M}_{\text{local}}$ token usage is 5,987 vs 2,093 tokens (2.9x). Overall, GIST trades additional computation on $\mathcal{M}_{\text{local}}$ for significantly stronger privacy protection before any information is exposed to $\mathcal{M}_{\text{remote}}$. External token costs remains comparable to PAPILLON as the anonymisation stage entirely runs on the trusted $\mathcal{M}_{\text{local}}$ rather than the chargeable $\mathcal{M}_{\text{remote}}$.