You Only Anonymize What Is Not Intent-Relevant: Suppressing Non-Intent Privacy Evidence

Anonymous ACL submission

Abstract

Anonymizing sensitive information in user text is essential for privacy, yet existing methods often apply uniform treatment across attributes, which can conflict with communicative intent and obscure necessary information. This is particularly problematic when personal attributes are integral to expressive or pragmatic goals. The central challenge lies in determining which attributes to protect, and to what extent, while preserving semantic and pragmatic functions. We propose INTENTANONY, a utility-preserving anonymization approach that performs intent-conditioned exposure control. INTENTANONY models pragmatic intent and constructs privacy inference evidence chains to capture how distributed cues support attribute inference. Conditioned on intent, it assigns each attribute an exposure budget and selectively suppresses non-intent inference pathways while preserving intent-relevant content, semantic structure, affective nuance, and interactional function. We evaluate INTENTANONY using privacy inference success rates, text utility metrics, and human evaluation. The results show an approximately 30% improvement in the overall privacy–utility trade-off, with notably stronger usability of anonymized text compared to prior state-of-the-art methods. Our code is available at `https://anonymous.4open.science/r/IA-S06`.

1 Introduction

Text anonymization aims to mitigate privacy risks in natural-language content while preserving the information necessary for faithful interpretation and effective use(Gadotti et al., 2024; Shahriar et al., 2025). This goal has become increasingly difficult in the era of large language models (LLMs), whose reasoning capabilities enable sensitive attribute inference even when explicit identifiers are removed(Patsakis and Lykousas, 2023; Wang et al., 2025; Wei et al., 2024; Harel et al., 2025). Recent work shows that privacy leakage often arises from distributed semantic, stylistic, and contextual cues rather than surface-level memorization(Li et al., 2024), rendering conventional masking-based anonymization insufficient in many settings (Staab et al., 2023; Sarkar et al., 2024).

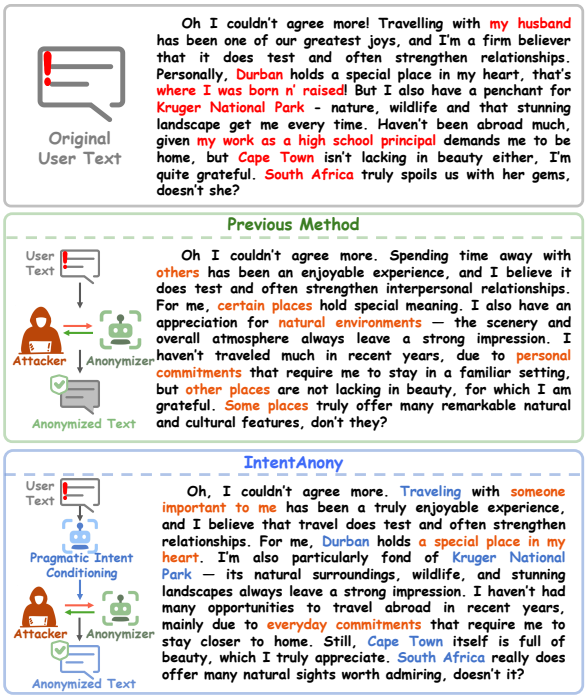

Figure 1: Anonymization under inference-based privacy threats. Compared with uniform rewriting methods that obscure sensitive details at the cost of communicative intent, INTENTANONY conditions anonymization on pragmatic intent and selectively suppresses non-intent privacy evidence, preserving semantic meaning and interactional function while reducing inference risk.

These inference-based privacy threats (Luo et al., 2025; Chen et al., 2025; Hui et al., 2024) have motivated a shift from simple de-identification toward semantics-aware anonymization (Kandpal et al., 2024; Du et al., 2025). Recent approaches formulate anonymization as a constrained transformation problem that balances privacy protection and

utility preservation through optimization, adversarial evaluation, or randomized rewriting strategies (Yang et al., 2025; Frikha et al., 2024; Kim et al., 2025). However, most existing methods apply privacy treatment in an attribute-agnostic manner, either uniformly masking detected attributes or broadly rewriting text without accounting for the communicative role that attributes play.

In practice, personal attributes vary in privacy sensitivity(Nissenbaum, 2009; Belen-Saglam et al., 2022). Some are deliberately disclosed to support communicative goals, while others seem benign in isolation but collectively enable inference. Uniform anonymization thus faces a trade-off: overly aggressive suppression harms intent-related utility, whereas insufficient suppression leaves inference channels exploitable by capable attackers Figure 1.

To address this limitation, we propose INTENTANONY, an anonymization approach that formulates privacy protection as an intent-conditioned exposure control problem. The central insight is that inference-based privacy risks emerge not from individual attributes alone, but from how distributed linguistic cues are pragmatically organized to support attribute inference under a given communicative goal. Accordingly, INTENTANONY explicitly models pragmatic intent and organizes privacy-relevant cues into privacy inference evidence chains, capturing how multiple textual spans jointly contribute to sensitive attribute inference. Conditioned on the recognized intent, each attribute is assigned an allowable exposure budget through an intent–attribute exposure matrix, which regulates the amount of inferential support retained in the anonymized text. This design enables selective suppression of non-intent inference pathways while preserving evidence that is functionally necessary for conveying meaning, stance, and interactional purpose. By intervening directly on inference-supporting evidence structures rather than surface identifiers, INTENTANONY reduces inference risk while maintaining semantic coherence and pragmatic intent.

We evaluate INTENTANONY using automatic privacy–utility metrics and human assessment, showing consistent reductions in inference risk together with improved preservation of semantic content, emotional tone, and communicative intent compared to strong masking and LLM-based rewriting baselines.

The main contributions of this work are summarized as follows:

- We introduce an intent-conditioned text anonymization approach that aligns privacy protection with communicative function by regulating personal attribute exposure.
- We propose a scene and intent conditioned exposure governance mechanism that enforces explicit attribute-level budgets for fine-grained, context-aware anonymization.
- Extensive privacy and utility evaluations, complemented by human evaluation, show that the proposed approach achieves a stronger privacy–utility trade-off, reducing attribute inference risk while preserving semantic coherence and communicative intent.

2 Related Work

Text Anonymization. Recent work has moved text anonymization beyond simple PII masking (Aahill, 2023) toward LLM-aware approaches that explicitly account for inference-based privacy risks. Many methods formulate anonymization as a constrained transformation problem that jointly optimizes privacy and utility, including iterative LLM-based frameworks with multi-component evaluation (Yang et al., 2025) and their analyses in personalized writing settings (Pasch and Cha, 2025; Manzanares-Salor and Sánchez, 2026). Complementary directions include conditional anonymization via private attribute randomization (Frikha et al., 2024), self-refining anonymizers based on adversarial distillation (Kim et al., 2025), and adversarial frameworks that leverage LLM inference itself for anonymization (Staab et al., 2025). Additional work explores inference-aware sanitization (Pilán et al., 2024), stylometric obfuscation (adv), context-preserving anonymization for structured or domain-specific text (Żarski and Janicki, 2025), and data-level anonymization for privacy-aware LLM deployment (Gardiner et al., 2024). Most existing methods are attribute-agnostic and overlook the communicative and pragmatic roles of personal attributes, which often results in unnecessary utility loss or insufficient disruption of inference pathways.

Privacy and Utility Trade-offs. Balancing privacy protection and textual utility is particularly challenging in the presence of inference-capable LLM adversaries. Prior work shows that modern LLMs can infer sensitive attributes from anonymized text even after explicit identifiers are removed (Staab et al., 2023). This has

motivated research on privacy–utility trade-offs in LLM-based anonymization. Recent methods suppress attribute inference through optimization- and evaluation-based strategies while preserving utility (Yang et al., 2025; Frikha et al., 2024), and through inference-aware sanitization techniques (Pilán et al., 2024; Manzanares-Salor and Sánchez, 2026). Parallel studies examine broader privacy risks and defenses for LLMs, including inference attacks and mitigation via differential privacy (Li et al., 2021; Shi et al., 2022) or privacy-preserving inference (Miranda et al., 2025; Yan et al., 2025). However, most existing methods lack intent-aware exposure control; in contrast, our work selectively suppresses non-intent inference cues while preserving intent-critical content.

3 Method

We present INTENTANONY, an intent-conditioned text anonymization approach that regulates personal attribute exposure according to communicative intent. Privacy leakage often arises from distributed linguistic cues whose inferential effect depends on the pragmatic role of attributes, causing uniform anonymization to either undercut inference resistance or unnecessarily distort meaning. INTENTANONY addresses this by formulating anonymization as intent-conditioned exposure control, explicitly aligning privacy protection with communicative function. Figure 2 provides an overview of the approach.

3.1 Pragmatic Intent Recognition

Given an input text x, the first step is to infer its underlying communicative intents. Instead of task-oriented intents commonly used in dialogue systems, we focus on pragmatic intents that characterize how language is functionally employed in social and expressive contexts, such as self-expression, social interaction, identity presentation, informational sharing, and sensitive disclosure. These intents govern not only what information is conveyed, but also which personal attributes are pragmatically relevant to meaning construction.

Formally, we define a finite set of pragmatic intents

$$\mathcal{I} = \{I_1, I_2, \ldots, I_K\}, \tag{1}$$

and cast intent recognition as a multi-label inference problem

$$f_{\text{intent}} : x \rightarrow \mathcal{I}(x), \tag{2}$$

where $\mathcal{I}(x) \subseteq \mathcal{I}$ denotes the set of intents expressed in x. Multiple intents may co-occur within a single text, reflecting the compositional nature of human communication.

We employ large language models as intent recognizers, exploiting their discourse-level reasoning to infer pragmatic intent from semantic and contextual evidence. The resulting intent set acts as a global constraint that guides subsequent anonymization decisions by distinguishing intent-relevant information from privacy-risk evidence.

3.2 Privacy Inference Evidence Chain

Sensitive personal attributes are rarely disclosed through isolated identifiers. Instead, they are often inferred from the aggregation of explicit, implicit, and contextual cues distributed across a text. Such compositional inference enables the recovery of latent attributes even in the absence of direct mentions, which limits the effectiveness of surface-level anonymization.

To capture this structure, we introduce privacy inference evidence chains. For each sensitive attribute $a \in \mathcal{A}$, an evidence chain is defined as

$$C_a = \{e_{a,1}, e_{a,2}, \ldots, e_{a,n}\}, \tag{3}$$

where each element $e_{a,i}$ denotes a textual cue contributing to the inference of a. These cues may be lexical, semantic, or contextual, and their inferential strength arises from their joint configuration rather than from any single element.

Each evidence element is assessed with respect to the recognized pragmatic intents of the text, enabling a distinction between intent-relevant evidence that supports communicative function and non-intent evidence that primarily amplifies inference pathways. Modeling privacy leakage at the level of evidence chains enables anonymization to intervene directly on inference-supporting structures, rather than relying on token-level masking or unstructured heuristic rewriting.

3.3 Scene–Intent Level Privacy Exposure Governance

To balance privacy protection and textual utility in an interpretable manner, we introduce an exposure governance mechanism conditioned on scene context and communicative intent. Rather than relying on static or attribute-agnostic anonymization rules, the mechanism determines how much information about each attribute may be retained according to

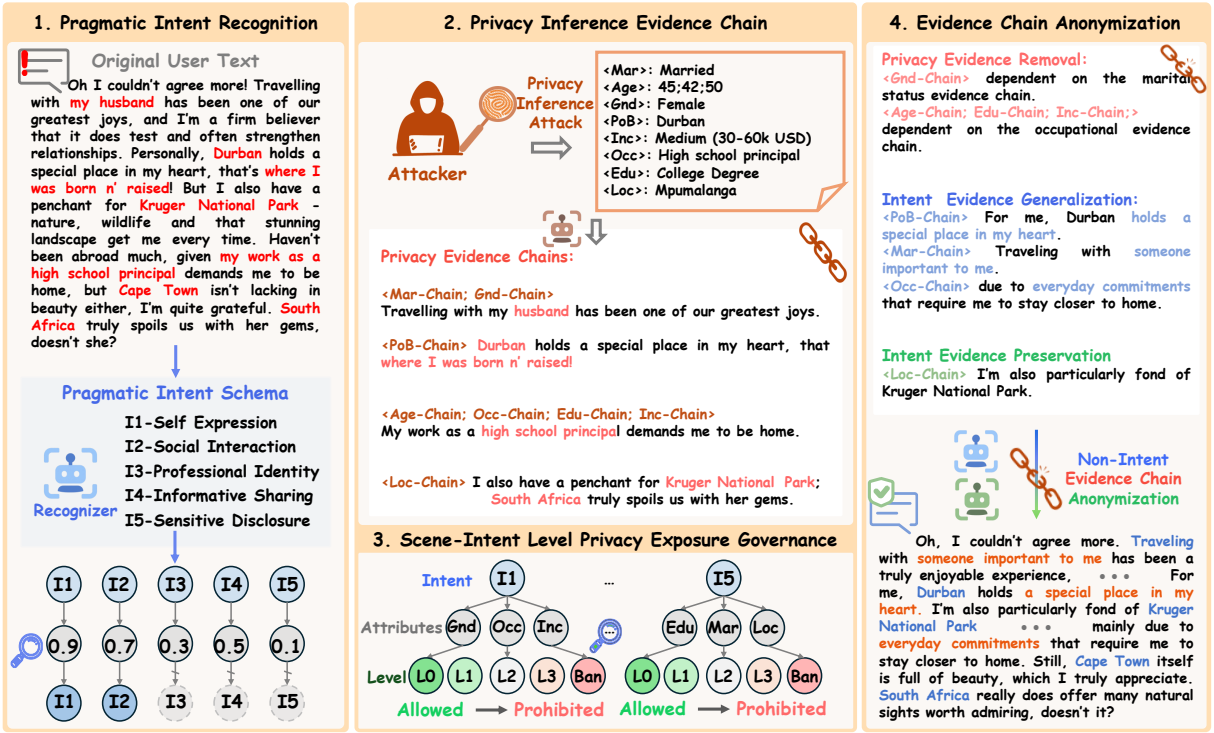

Figure 2: Overview of the proposed intent-aware anonymization framework. The pipeline includes four stages: (1) pragmatic intent recognition to identify communicative intents in the input text; (2) privacy inference evidence chain construction that organizes sensitive attributes into intent-grounded evidence chains; (3) scene–intent level privacy exposure governance for determining appropriate anonymization levels; and (4) evidence chain anonymization, which selectively rewrites or removes non-intent evidence while preserving intent-relevant content.

the communicative intent expressed in the text and its surrounding context.

We define an ordered set of privacy exposure granularity levels

$$\mathcal{E} = \{L_0, L_1, L_2, L_3, \text{BAN}\}, \tag{4}$$

where increasing levels impose stricter exposure constraints, ranging from minimal modification to complete suppression.

Given a scene representation $s(x)$ inferred from context, we define an exposure governance function

$$G : \mathcal{S} \times \mathcal{I} \times \mathcal{A} \rightarrow \mathcal{E}, \tag{5}$$

which assigns each combination of scene, intent, and attribute a maximum allowable exposure level. This mapping reflects how the appropriateness of attribute disclosure varies across different communicative contexts.

When a text expresses multiple intents, we adopt a conservative aggregation strategy and assign each attribute a an effective exposure budget

$$\ell_a = \min_{I_k \in \mathcal{I}(x)} G\big(s(x), I_k, a\big). \tag{6}$$

The resulting budget constrains the total contribution of the corresponding evidence chain C_a in the anonymized output. By conditioning exposure control on both scene context and communicative intent, this mechanism enables fine-grained anonymization while avoiding insufficient protection and unnecessary removal of intent-relevant information.

3.4 Evidence Chain Anonymization

Given the intent-conditioned exposure budgets, IntentAnony generates an anonymized text $\tilde{x}$ through evidence-chain-guided rewriting. Anonymization operates at the level of privacy inference evidence chains rather than isolated tokens, enabling direct intervention on the structured aggregation of cues that supports attribute inference.

For each sensitive attribute $a \in \mathcal{A}$ with evidence chain C_a, we assess the functional role of individual evidence elements with respect to the recognized communicative intents $\mathcal{I}(x)$. Based on this assessment, the evidence chain C_a is conceptually decomposed into two complementary subsets: an intent-relevant component C_a^{intent}, which is necessary for realizing the communicative intent, and a non-intent component $C_a^{\text{non-intent}}$, which primarily strengthens sensitive attribute inference without

contributing to the intended meaning. This decomposition allows intent-relevant evidence to be preserved or generalized, while non-intent evidence is selectively attenuated to reduce inference-based privacy risk.

Anonymization is governed by the exposure budget ℓ_a associated with attribute a, which constrains the expected inference risk induced by retained evidence. Specifically, the anonymized text $\tilde{x}$ is required to satisfy

$$\mathbb{E}\big[\,\mathcal{R}(a \mid \tilde{x})\,\big] \;\leq\; \ell_a, \tag{7}$$

where $\mathcal{R}(a \mid \tilde{x})$ denotes the inference risk of attribute a given the text, as estimated by privacy-oriented inference prompts, and the expectation is taken over the stochastic generation process of the language model. To meet this constraint, evidence in $C_a^{\text{non-intent}}$ is suppressed or removed, while evidence in C_a^{intent} is preserved through abstraction or generalization when necessary. This process is implemented via constrained LLM-based rewriting, in which prompts explicitly encode intent requirements together with attribute-specific exposure limits. By intervening at the level of evidence chains and explicitly constraining inference risk, INTENTANONY disrupts compositional inference pathways while preserving semantic coherence and interactional quality.

4 Experimental Set-up

Datasets. We evaluate our method on two datasets covering realistic and controlled settings. Both datasets consist of Reddit-style text annotated with personal attributes and exhibit properties comparable to authentic user-generated content. *PersonalReddit*(Staab et al., 2023) contains Reddit-style Q&A pairs with naturally embedded attributes, while *SynthPAI* (Yukhymenko et al., 2024) consists of Reddit-style comments with systematically varied attribute combinations. Additional data preprocessing details are provided in Appendix C.1.

Evaluation Metrics. Utility is evaluated along four complementary dimensions: readability, hallucination control, semantic preservation, and surface-level similarity. Readability and hallucination capture textual fluency and faithfulness, while semantic preservation assesses retention of the original meaning and intent; BLEU and ROUGE quantify lexical overlap. Privacy protection is evaluated via attribute inference attacks by a strong LLM adversary, where attack accuracy indicates residual privacy leakage. We jointly report privacy and utility to assess whether anonymization reduces inference risk while preserving communicative value.

Models in Comparison. We compare INTENTANONY with representative anonymization systems covering commercial solutions, rule-based pipelines, and LLM-driven approaches.

- AZURE (Aahill, 2023) Text Anonymization is a commercial PII masking system that removes explicit identifiers using fixed entity detectors.
- DIPPER (Krishna et al., 2023) is a paraphrasing-based baseline that rewrites text while preserving overall semantics.
- ADV. ANON. (Staab et al., 2025) is an LLM-based anonymizer that leverages adversarial feedback to reduce attribute inference risk.
- RUPTA (Yang et al., 2025) is a utility-oriented anonymization framework that jointly considers privacy risk reduction and semantic consistency.

Implementation Details. All main experiments are conducted using DeepSeek-V3.2 (DeepSeek-AI et al., 2025). Following prior work, the LLM-based baselines ADV. ANON. and RUPTA are implemented on the same model to ensure fair comparison. To examine model-agnostic performance, INTENTANONY is additionally evaluated on GLM-4.7 (Zeng et al., 2025; AI, 2025), GPT-5.2 (OpenAI, 2025), and Gemini-3-Pro (DeepMind, 2025). The original text is included for reference, with further details provided in Appendix C.2.

5 Experimental Results

5.1 Overall Performance

We conduct a unified evaluation of anonymization methods by jointly examining privacy leakage, text utility, and their aggregated performance on two datasets with distinct attribute distributions. Privacy is assessed via attribute inference accuracy across multiple sensitive dimensions, while utility is measured using complementary indicators of readability, hallucination control, and lexical similarity. This evaluation setting enables a direct comparison of how different methods balance inference resistance against linguistic preservation. As reported in Table 1, with all results averaged over 5 independent runs, INTENTANONY achieves the strongest overall performance among all evaluated methods. Across both datasets, it consistently reduces inference accuracy on sensitive attributes relative to paraphrasing-based approaches,

Metric	PersonalReddit						SynthPAI					
	Orig.	**Azure**	**Dipper**	**A.A.**	**RUPTA**	**Ours**	**Orig.**	**Azure**	**Dipper**	**A.A.**	**RUPTA**	**Ours**
Age	0.500	0.379	0.482	0.329	0.379	0.400	0.387	0.342	0.385	0.325	0.333	0.333
Edu	0.765	0.643	0.724	0.520	0.735	0.541	0.690	0.563	0.688	0.344	0.469	0.375
Gnd	0.890	0.787	0.856	0.748	0.780	0.732	0.943	0.829	0.854	0.756	0.707	0.780
Inc	0.736	0.755	0.712	0.679	0.651	0.745	0.745	0.658	0.632	0.539	0.632	0.605
Loc	0.593	0.102	0.561	0.163	0.045	0.098	0.328	0.103	0.362	0.000	0.138	0.138
Mar	0.757	0.486	0.806	0.446	0.365	0.338	0.833	0.660	0.766	0.426	0.596	0.489
Occ	0.574	0.347	0.512	0.231	0.546	0.204	0.707	0.634	0.704	0.127	0.634	0.338
PoB	0.549	0.155	0.451	0.099	0.014	0.099	0.333	0.000	0.429	0.000	0.143	0.000
Privacy ↓	0.650	0.411	0.614	0.365	0.417	**0.353**	0.607	0.499	0.579	**0.334**	0.474	0.410
Mean	1.000	0.787	0.937	0.833	0.826	0.956	1.000	0.727	0.854	0.847	0.920	0.981
Read	1.000	0.371	0.970	0.998	0.990	1.000	1.000	0.291	0.827	0.988	0.862	0.946
Hall	1.000	1.000	0.948	0.991	0.873	1.000	1.000	1.000	0.834	0.995	0.961	0.998
BLEU	1.000	0.824	0.247	0.622	0.745	0.852	1.000	0.798	0.130	0.490	0.712	0.849
ROUGE	1.000	0.957	0.676	0.804	0.878	0.930	1.000	0.952	0.617	0.730	0.912	0.944
Utility ↑	1.000	0.833	0.625	0.789	0.840	**0.923**	1.000	0.807	0.528	0.721	0.846	**0.923**
Overall ↑	-	0.201	-0.320	0.227	0.198	**0.379**	-	-0.015	-0.426	0.171	0.065	**0.247**

Table 1: Unified performance comparison of anonymization methods on the PersonalReddit and SynthPAI datasets. A.A. denotes Adv. Anon.; best and second-best results are shown in bold and underlined, respectively.

while avoiding the substantial utility degradation observed in more aggressive anonymization strategies. INTENTANONY maintains high scores on utility-related metrics, indicating that privacy improvements are not obtained at the expense of semantic coherence or surface-level fidelity.

The observed performance trends remain stable across datasets constructed under different generation procedures, suggesting that the proposed intent-conditioned exposure control is not tightly coupled to a specific data distribution. Overall, these results indicate that incorporating communicative intent into anonymization decisions yields a more reliable balance between privacy protection and text utility than existing baselines.

5.2 Robustness Across LLM

To assess the robustness of INTENTANONY across different backbone language models, we evaluate its performance using four representative LLMs on both datasets. Table 2 reports privacy, utility, and overall scores under identical anonymization settings. Across all backbones, INTENTANONY consistently achieves strong privacy protection while preserving high text utility, resulting in stable overall performance. Notably, stronger backbones such as GPT-5.2 and Gemini-3-Pro yield particularly favorable trade-offs, achieving lower privacy leakage and higher utility scores compared to lighter models. This suggests that INTENTANONY can effectively leverage the reasoning and generation capabilities of more advanced LLMs without relying on model-specific behaviors. Overall, the consistent performance trends across diverse backbones

LLM	GLM-4.7	DS-V3.2	GPT-5.2	Gemini-3-Pro
Personal Reddit				
Privacy ↓	0.370	0.353	0.379	**0.345**
Utility ↑	0.939	0.923	**0.946**	0.916
Overall ↑	0.370	0.379	0.363	**0.385**
SynthPAI				
Privacy ↓	0.425	0.410	**0.403**	**0.403**
Utility ↑	0.939	0.923	**0.953**	0.919
Overall ↑	0.238	0.247	**0.289**	0.255

Table 2: Performance of different LLM backbones on the Personal Reddit and SynthPAI datasets. DS-V3.2 denotes DeepSeek-V3.2; best results are shown in bold.

indicate that intent-conditioned exposure control generalizes well and remains effective under heterogeneous deployment settings.

5.3 Privacy–Utility Trade-off

To analyze the privacy–utility trade-off, we evaluate anonymized outputs under five privacy granularity levels, ranging from L0, L1, L2, and L3 to BAN, which correspond to progressively stricter anonymization constraints. As the privacy level increases, inference-based attack success is expected to decrease, while text utility may gradually degrade. Experiments are conducted across multiple commercial large language models to examine the stability of this trade-off under different inference behaviors and generation characteristics.

Figure 3 shows the relationship between inference accuracy and mean text utility across different privacy levels. As privacy strength increases from L0 to BAN, inference accuracy consistently decreases, indicating improved resistance to infer-

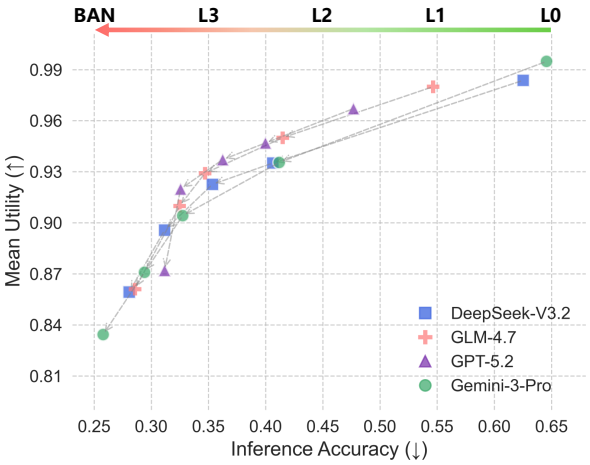

Figure 3: Privacy–utility trade-off across five privacy granularity levels (L0 to BAN), showing the relationship between inference attack accuracy and mean text utility on multiple commercial language models.

ence attacks, accompanied by a gradual decline in text utility. Importantly, similar trajectories are observed across all evaluated models, suggesting that the privacy–utility trade-off is robust to model choice. At intermediate privacy levels (e.g., L1 and L2), anonymized texts retain relatively high utility while substantially limiting inference accuracy, representing a favorable operating region. These results demonstrate that the proposed framework supports fine-grained control over privacy strength, enabling practitioners to balance privacy and utility according to task-specific requirements.

5.4 Semantic Similarity Distribution

Semantic similarity between anonymized texts and their originals reflects how well semantic content is preserved during anonymization. Examining the distribution of similarity scores, rather than only their averages, helps reveal per-sample variability in semantic preservation and the stability of anonymization behavior. Figure 4 compares the semantic similarity distributions of INTENTANONY with representative baselines. INTENTANONY exhibits distributions more concentrated toward higher similarity values, with noticeably less mass in low-similarity regions. By contrast, baseline methods show broader distributions with heavier lower tails, indicating more frequent semantic deviation. These results indicate that INTENTANONY preserves semantic content more consistently and better maintains communicative intent during anonymization. Additional distributional results on the SynthPAI dataset are provided in Appendix A.1.

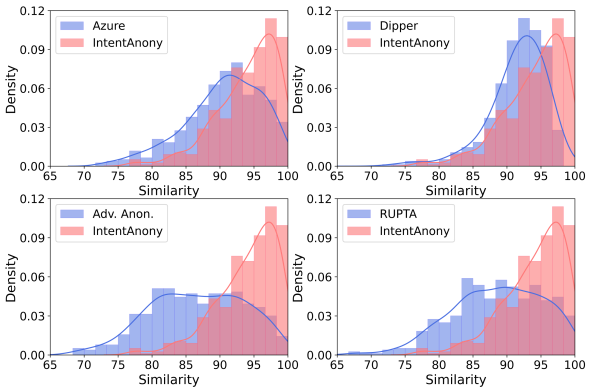

Figure 4: Distribution of semantic similarity scores between anonymized and original texts on the PersonalReddit dataset. INTENTANONY yields distributions more concentrated toward higher values than baseline methods, indicating more consistent preservation of original semantics and communicative intent.

5.5 Intent Change

Preserving communicative intent is a critical yet often overlooked requirement in text anonymization, as intent drift can undermine functional equivalence and downstream usability. Figure 5 compares intent preservation across anonymization methods and reveals substantial variation among rewriting-based approaches. DIPPER achieves relatively high intent overlap but lower stability, suggesting that paraphrasing may introduce intent fluctuations, while AZURE and ADV. ANON. exhibit moderate trade-offs between anonymization strength and intent preservation.

RUPTA improves intent consistency through constrained rewriting, though measurable drift remains. In contrast, INTENTANONY attains the highest scores on both Intent Overlap and Stability F1, indicating minimal deviation from the original intent. Overall, these results show that explicit intent

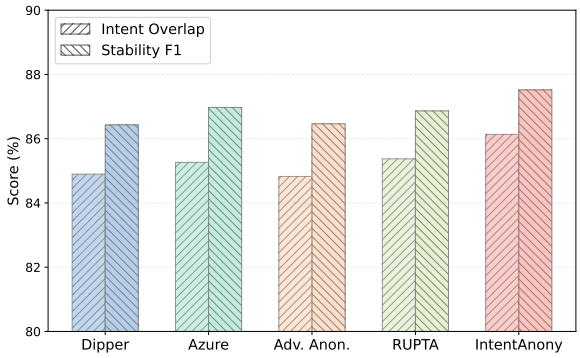

Figure 5: Comparison of intent preservation across different anonymization methods, measured by Intent Overlap and Stability F1, where higher scores indicate better alignment with the original communicative intent.

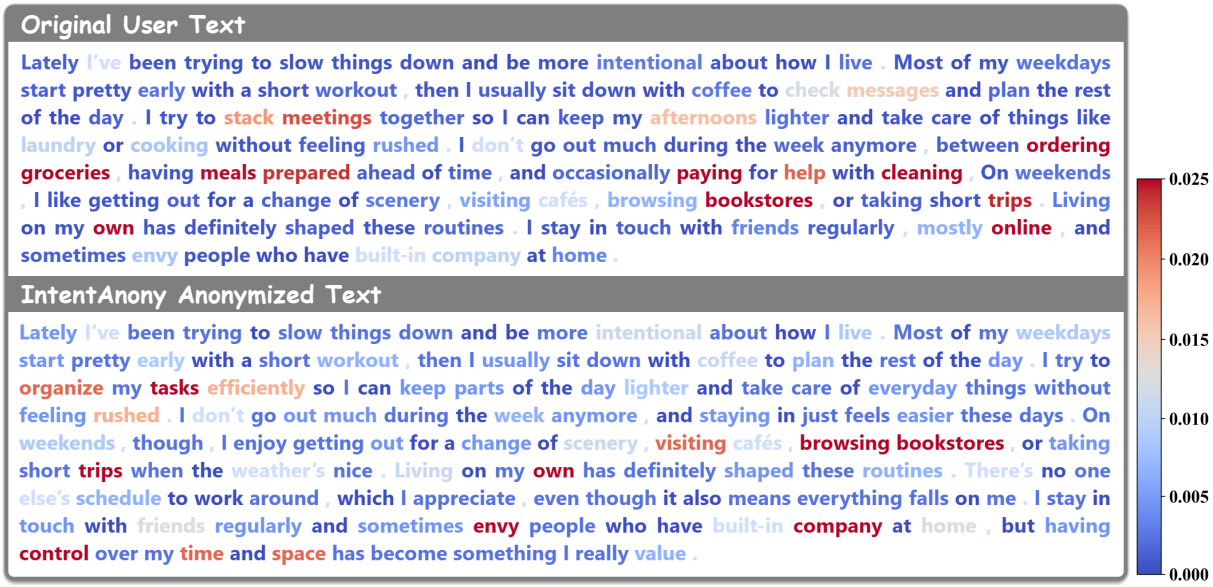

Figure 6: Token-level visualization of privacy contribution scores for the original text (top) and the INTENTANONY anonymized text (bottom). Colors indicate the relative contribution of tokens to sensitive attribute inference. Non-intent privacy cues with high contribution in the original text are selectively attenuated after intent-conditioned anonymization, while intent-relevant semantic content is largely preserved.

modeling is essential for stable intent preservation under anonymization.

5.6 Token-Level Privacy Inference Analysis

Figure 6 illustrates token-level privacy contribution scores before and after intent-conditioned anonymization, highlighting how INTENTANONY mitigates inference-based privacy risks while preserving communicative intent. Following prior findings that attribute inference relies on structured contextual cues rather than isolated tokens (Ren et al., 2024; Zheng et al., 2025), we estimate each token's contribution to sensitive attribute inference using privacy-oriented inference prompts, yielding scores that capture functional inference support instead of raw attention weights.

In the original text, tokens associated with lifestyle routines, social habits, and living arrangements exhibit elevated contribution scores, indicating that privacy leakage often arises from the accumulation of semantically meaningful but non-explicit cues. Under intent-conditioned anonymization, INTENTANONY selectively suppresses or abstracts such non-intent privacy cues, resulting in a clear attenuation of high-contribution tokens, while preserving tokens essential for expressing reflective intent and narrative coherence. This analysis shows that INTENTANONY disrupts inference-supporting evidence structures at the token level, rather than relying on uniform masking or indiscriminate rewriting, thereby reducing privacy risk without compromising semantic fidelity or communicative intent.

5.7 Human Evaluation

We conduct a human evaluation using a custom-built interactive annotation system to assess anonymization quality beyond automatic metrics, with a focus on privacy protection, semantic and intent fidelity, and social acceptability. Detailed descriptions of the evaluation interface, evaluation protocol, and score aggregation procedures are provided in Appendix B.

6 Conclusion

We presented INTENTANONY, an intent-conditioned approach to text anonymization that regulates attribute exposure to mitigate inference-based privacy risks while preserving communicative utility. By replacing uniform masking or unconstrained rewriting with exposure budgets derived from pragmatic intent, INTENTANONY selectively attenuates non-intent evidence chains and retains intent-relevant content, resulting in anonymized text that better maintains semantics, affective nuance, and interactional coherence. Extensive experiments using automatic metrics and human evaluation show that INTENTANONY achieves a consistently improved privacy–utility balance and reduces sensitive attribute inference across datasets and backbone language models.

Limitations

While INTENTANONY provides an effective approach for mitigating inference-based privacy risks under intent constraints, several aspects merit further consideration.

INTENTANONY builds on pragmatic intent recognition to guide exposure governance and evidence chain anonymization. In most cases, contemporary large language models offer sufficiently reliable intent understanding to support this process. Nevertheless, communicative intents can be nuanced, overlapping, or implicitly expressed, suggesting that more refined intent modeling or uncertainty-aware intent representations could further enhance robustness in complex discourse settings. The formulation of privacy inference evidence chains is grounded in the inference behavior of large language models, which enables realistic simulation of modern privacy threats. At the same time, inference cues may vary across model families or future model iterations. While our results demonstrate consistent trends across multiple strong backbones, extending evidence chain modeling to account for broader or evolving inference behaviors remains a promising direction for future work. Finally, INTENTANONY is designed as a practical anonymization method that operates at the level of textual rewriting under inference-based threats. It does not aim to provide formal worst-case privacy guarantees, but rather complements existing formal privacy mechanisms by offering fine-grained, intent-aware control over semantic exposure. Integrating intent-conditioned anonymization with provable privacy guarantees constitutes an interesting avenue for future research.

Ethical Considerations

This paper studies text anonymization under inference-based privacy threats, with the goal of improving privacy protection while preserving communicative intent and textual utility. We acknowledge that anonymization technologies can serve both protective and potentially harmful purposes if misapplied. To promote responsible use, we emphasize transparency in our modeling assumptions, design choices, and limitations, which are discussed throughout the paper. Our approach is developed as a defensive mechanism against attribute inference and profiling, and is not intended to facilitate surveillance, re-identification, or misuse of personal information. All experiments are conducted on existing benchmark datasets and publicly accessible textual resources. We view this work as a contribution toward more responsible and intent-aware deployment of language technologies, supporting privacy protection without undermining expressive autonomy. We emphasize that the proposed method is not a substitute for legal, regulatory, or formally provable privacy guarantees.

References

Adversarial stylometry. Wikipedia, Accessed 2025.

Aahill. 2023. What is azure ai language - azure ai services. https://learn.microsoft.com/en-us/azure/ai-services/language-service/overview. Accessed on Jan 12, 2024.

Zhipu AI. 2025. Glm-4.7: Advancing the coding capability. https://z.ai/blog/glm-4.7. Accessed on Dec 22, 2025.

Rahime Belen-Saglam, Jason RC Nurse, and Duncan Hodges. 2022. An investigation into the sensitivity of personal information and implications for disclosure: a uk perspective. *Frontiers in Computer Science*, 4:908245.

Tiejin Chen, Pingzhi Li, Kaixiong Zhou, Tianlong Chen, and Hua Wei. 2025. Unveiling privacy risks in multimodal large language models: Task-specific vulnerabilities and mitigation challenges. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4573–4586, Vienna, Austria. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 pro. https://deepmind.google/models/gemini/pro/. Accessed on Nov 18, 2025.

Aixin Liu DeepSeek-AI, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Yuntao Du, Zitao Li, Ninghui Li, and Bolin Ding. 2025. Beyond data privacy: New privacy risks for large language models. *arXiv preprint arXiv:2509.14278*.

Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. Incognitext: Privacy-enhancing conditional text anonymization via llm-based private attribute randomization. *arXiv preprint arXiv:2407.02956*.

Andrea Gadotti, Luc Rocher, Florimond Houssiau, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. 2024. Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*, 10(29):eadn7053.

Shayna Gardiner, Tania Habib, Kevin Humphreys, Masha Azizi, Frederic Mailhot, Anne Paling, Preston Thomas, and Nathan Zhang. 2024. Data anonymization for privacy-preserving large language model fine-tuning on call transcripts. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 64–75.

Re'em Harel, Niv Gilboa, and Yuval Pinter. 2025. Token-level privacy in large language models. *arXiv preprint arXiv:2503.03652*.

Bo Hui, Haolin Yuan, Neil Gong, Philippe Burlina, and Yinzhi Cao. 2024. Pleak: Prompt leaking attacks against large language model applications. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3600–3614.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A. Choquette-Choo, and Zheng Xu. 2024. User inference attacks on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18238–18265, Miami, Florida, USA. Association for Computational Linguistics.

Kyuyoung Kim, Hyunjun Jeon, and Jinwoo Shin. 2025. Self-refining language model anonymizers via adversarial distillation. *Preprint*, arXiv:2506.01420.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36:27469–27500.

Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, Bo Li, Bingsheng He, and Dawn Song. 2024. Llm-pbe: Assessing data privacy in large language models. *Proc. VLDB Endow.*, 17(11):3201–3214.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*.

Xinjian Luo, Ting Yu, and Xiaokui Xiao. 2025. Prompt inference attack on distributed large language model inference frameworks. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, CCS '25, page 1739–1753, New York, NY, USA. Association for Computing Machinery.

Benet Manzanares-Salor and David Sánchez. 2026. A comparative analysis, enhancement and evaluation of text anonymization with pre-trained large language models. *Expert Systems with Applications*, 297:129474.

Michele Miranda, Elena Sofia Ruzzetti, Andrea Santilli, Fabio Massimo Zanzotto, Sébastien Bratières, and Emanuele Rodolà. 2025. Preserving privacy in large language models: A survey on current threats and solutions. *Transactions on Machine Learning Research*.

Helen Nissenbaum. 2009. Privacy in context: Technology, policy, and the integrity of social life. In *Privacy in context*. Stanford University Press.

OpenAI. 2025. Introducing gpt-5.2. `https://openai.com/index/introducing-gpt-5-2/`. Accessed on Dec 11, 2025.

Stefan Pasch and Min Chul Cha. 2025. Balancing privacy and utility in personal LLM writing tasks: An automated pipeline for evaluating anonymizations. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 32–41, Albuquerque, New Mexico. Association for Computational Linguistics.

Constantinos Patsakis and Nikolaos Lykousas. 2023. Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports*, 13(1):16026.

Ildikó Pilán, Benet Manzanares-Salor, David Sánchez, and Pierre Lison. 2024. Truthful text sanitization guided by inference attacks. *arXiv preprint arXiv:2412.12928*.

Jie Ren, Qipeng Guo, Hang Yan, Dongrui Liu, Quanshi Zhang, Xipeng Qiu, and Dahua Lin. 2024. Identifying semantic induction heads to understand in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6916–6932, Bangkok, Thailand. Association for Computational Linguistics.

Atiquer Rahman Sarkar, Yao-Shun Chuang, Noman Mohammed, and Xiaoqian Jiang. 2024. De-identification is not always enough. *Preprint*, arXiv:2402.00179.

Sakib Shahriar, Rozita Dara, and Rajen Akalu. 2025. A comprehensive review of current trends, challenges, and opportunities in text data privacy. *Computers & Security*, 151:104358.

Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. 2022. Selective differential privacy for language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2848–2859, Seattle, United States. Association for Computational Linguistics.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations (ICLR)*, pages 1–17.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2025. Language models are advanced anonymizers. In *International Conference on Learning Representations (ICLR)*.

Shang Wang, Tianqing Zhu, Bo Liu, Ming Ding, Dayong Ye, Wanlei Zhou, and Philip Yu. 2025. Unique security and privacy threats of large language models: A comprehensive survey. *ACM Comput. Surv.*, 58(4).

Jiankun Wei, Abdulrahman Abdulrazzag, Tianchen Zhang, Adel Muursepp, and Gururaj Saileshwar. 2024. Privacy risks of speculative decoding in large language models. *arXiv preprint arXiv:2411.01076*.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2025. On protecting the data privacy of large language models (llms) and llm agents: A literature review. *High-Confidence Computing*, 5(2):100300.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2025. Robust utility-preserving text anonymization based on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 28922–28941. Association for Computational Linguistics.

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

Tymon Lesław Żarski and Artur Janicki. 2025. Enhancing privacy while preserving context in text transformations by large language models. *Information*, 16(1):49.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025. Attention heads of large language models. *Patterns*, 6(2).

Appendix

A Additional Analysis **13**
- A.1 Semantic Similarity Distribution on the SynthPAI Dataset 13
- A.2 Intent Robustness under Anonymization 13
- A.3 Pricing Cost Analysis 13

B Human Evaluation **14**
- B.1 Human Subjects and Evaluation Procedure. 14
- B.2 Human Evaluation Protocol 14
- B.3 Human Evaluation System 15

C Experimental Details **15**
- C.1 Dataset Details 15
- C.2 Implementation Details 16

D Prompts **17**
- D.1 Pragmatic Intent Recognition . . . 17
- D.2 Personal Attribute Privacy Inference 17
- D.3 Privacy Inference Evidence Chain 18
- D.4 Pragmatic Intent Anonymization . 19
- D.5 Utility Evaluation 20
- D.6 Inference Validation 21

E Anonymization Examples **21**

A Additional Analysis

A.1 Semantic Similarity Distribution on the SynthPAI Dataset

To complement the analysis on the Personal Reddit dataset presented in the main paper, we further investigate semantic similarity distributions on the SynthPAI dataset, which provides controlled and systematically varied attribute configurations. Compared with naturally occurring user text, SynthPAI allows for more precise examination of anonymization behavior under diverse yet structured semantic and contextual settings, offering an additional perspective on method robustness. As shown in Figure 7, INTENTANONY consistently produces semantic similarity distributions that are more concentrated toward higher values than those of baseline methods, accompanied by noticeably reduced probability mass in lower-similarity regions. This distributional pattern suggests that INTENTANONY introduces fewer large semantic deviations across samples, yielding more stable preservation of original meaning and communicative intent. The consistency of these trends across controlled data conditions indicates that the intent-conditioned anonymization mechanism generalizes beyond naturally occurring text and is not overly dependent on dataset-specific characteristics.

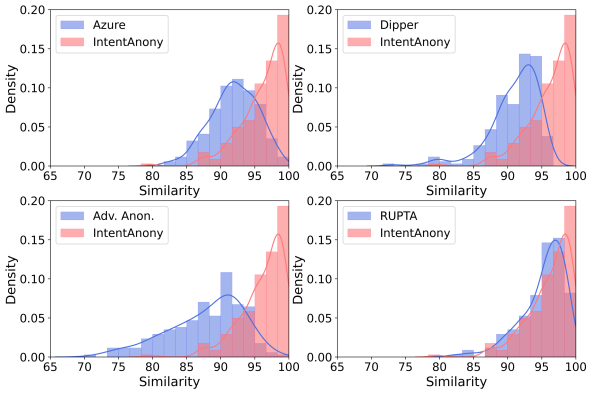

Figure 7: Distribution of semantic similarity scores between anonymized and original texts on the SynthPAI dataset. Compared with baseline methods, INTENTANONY yields distributions more concentrated toward higher similarity values, indicating more consistent preservation of semantic content and communicative intent under controlled attribute settings.

A.2 Intent Robustness under Anonymization

Intent recognition robustness is a key factor in intent-aware anonymization, as errors at this stage may propagate to downstream processing. Rather than assuming uniform intent recognition capability across models, we analyze how reliably different large language models align with manually annotated communicative intents. As shown in Table 3, the results reveal clear performance differences across models: Gemini-3-Pro achieves the strongest alignment with ground-truth intents, as reflected by higher NDCG@2 and J-Acc scores, while GPT-5.2 and GLM-4.7 show comparable performance and DeepSeek-V3.2 lags behind. These observations indicate that intent recognition quality varies substantially across contemporary language models, and that model choice can materially affect the reliability of intent-aware anonymization frameworks that rely on semantic intent as an intermediate representation.

Method	NDCG@2 ↑	J-Acc ↑	P ↓	U ↑
GLM-4.7	0.825	0.730	0.370	0.939
DeepSeek-V3.2	0.807	0.697	0.379	0.926
GPT-5.2	0.836	0.752	0.379	**0.946**
Gemini-3-Pro	**0.863**	**0.782**	**0.345**	0.916

Table 3: Intent recognition robustness of different large language models on manually annotated texts. NDCG@2 and J-Acc (Jaccard Accuracy) measure intent alignment, while P and U denote inference privacy accuracy and utility, respectively.

A.3 Pricing Cost Analysis

Beyond privacy protection and text utility, the monetary cost associated with large language model usage is an important practical factor for deploying anonymization methods at scale. We therefore analyze the pricing overhead of different approaches in terms of their token consumption. The cost is measured by aggregating the total number of input and output tokens consumed across all model calls required to anonymize a single sample. To ensure a fair comparison across methods with different prompting strategies and processing pipelines, all costs are reported as relative values normalized to INTENTANONY, which is set to $1.0\times$.

The resulting cost comparison is summarized in Table 4. Methods relying on multi-round refinement or iterative evaluation incur substantially higher pricing overhead due to repeated model invocations for adversarial inference, utility assessment, and rewrite refinement. For example, ADV. ANON. alternates between adversarial generation and evaluation, while RUPTA introduces additional calls for iterative privacy–utility balancing, both of which significantly increase token con-

sumption. In contrast, INTENTANONY employs intent-conditioned exposure control with evidence-chain analysis, allowing anonymization to be completed in a single generation pass without iterative feedback. This design markedly reduces model invocations and token usage, yielding a more favorable balance between inference resistance and practical efficiency.

Method	Anonymization Strategy	Relative Cost
ADV. ANON.	Adversarial multi-round refinement	2.8×
RUPTA	Iterative privacy–utility evaluation	2.2×
OURS	Intent-conditioned exposure control	1.0×

Table 4: Relative token consumption and pricing comparison of anonymization methods. Costs are normalized by INTENTANONY (ours), which is set to 1.0×.

B Human Evaluation

B.1 Human Subjects and Evaluation Procedure.

This study includes a human evaluation component, in which human evaluators assess the quality of anonymized texts. Prior to participation, all participants were provided with clear and comprehensive written instructions describing (i) the purpose and overall procedure of the evaluation, (ii) the evaluation dimensions and corresponding rating criteria, (iii) the fact that the evaluation does not involve exposure to real identities or sensitive personal information, and (iv) the voluntary nature of participation, with the option to withdraw at any stage without penalty.

Participants were exposed only to anonymized text content. The evaluation task poses no psychological, legal, or economic risks, and no known potential harm to participants is involved. All instructions were delivered in written form, and informed consent was obtained from all participants before the evaluation began.

B.2 Human Evaluation Protocol

We conduct a human evaluation to assess anonymization quality beyond automatic and surface-level metrics. The evaluation focuses on three complementary dimensions: *Perceived Privacy Protection* (PPP), *Semantic and Intent Fidelity* (SIF), and *Social Acceptability and Expressiveness* (SAE). In addition, we report the *Anonymization User Preference Index* (AUPI) as an aggregated preference-based indicator that jointly reflects privacy adequacy and text usability.

The evaluation is performed on two benchmark datasets, with 100 randomly sampled instances from each. For every instance, anonymized outputs are generated by AZURE, DIPPER, ADV. ANON., RUPTA, and our proposed method, INTENTANONY. Ten independent human evaluators participate in the study. For each sample, participants are presented with the original text together with anonymized outputs from different methods, and are asked to rate each anonymized version on a 1–10 Likert scale for PPP, SIF, and SAE. Method identities are concealed and output orders are randomized to minimize potential bias. Final scores are obtained by averaging ratings across human evaluators and samples. AUPI is computed by aggregating the ratings across all evaluation dimensions, yielding an overall preference score that reflects participants' holistic judgments of anonymization quality.

Table 5 summarizes the human evaluation results. Azure ACHIEVES attains relatively high PPP scores, but exhibits substantially lower SIF and SAE scores, indicating that aggressive masking often compromises semantic fidelity and social naturalness. DIPPER better preserves surface semantics but provides weaker privacy protection, suggesting that paraphrasing alone is insufficient against inference-based leakage. RUPTA and ADV. ANON. achieve a more balanced trade-off through rewriting-based anonymization. In contrast, INTENTANONY consistently achieves the highest or near-highest scores across all three dimensions and obtains the best overall mean score, demonstrating a stronger balance between privacy protection, semantic preservation, and social acceptability.

Method	PPP	SIF	SAE	Mean
Azure	6.50	4.22	2.27	4.33
Dipper	4.43	6.94	6.51	5.95
Adv. Anon.	**7.50**	6.74	6.82	7.02
RUPTA	6.30	6.45	6.52	6.42
IntentAnony	7.48	**7.53**	**7.96**	**7.66**

Table 5: Human evaluation of anonymization methods across perceived privacy protection (PPP), semantic & intent fidelity (SIF), and social acceptability & expressiveness (SAE). Higher scores indicate better performance. Best results are shown in bold.

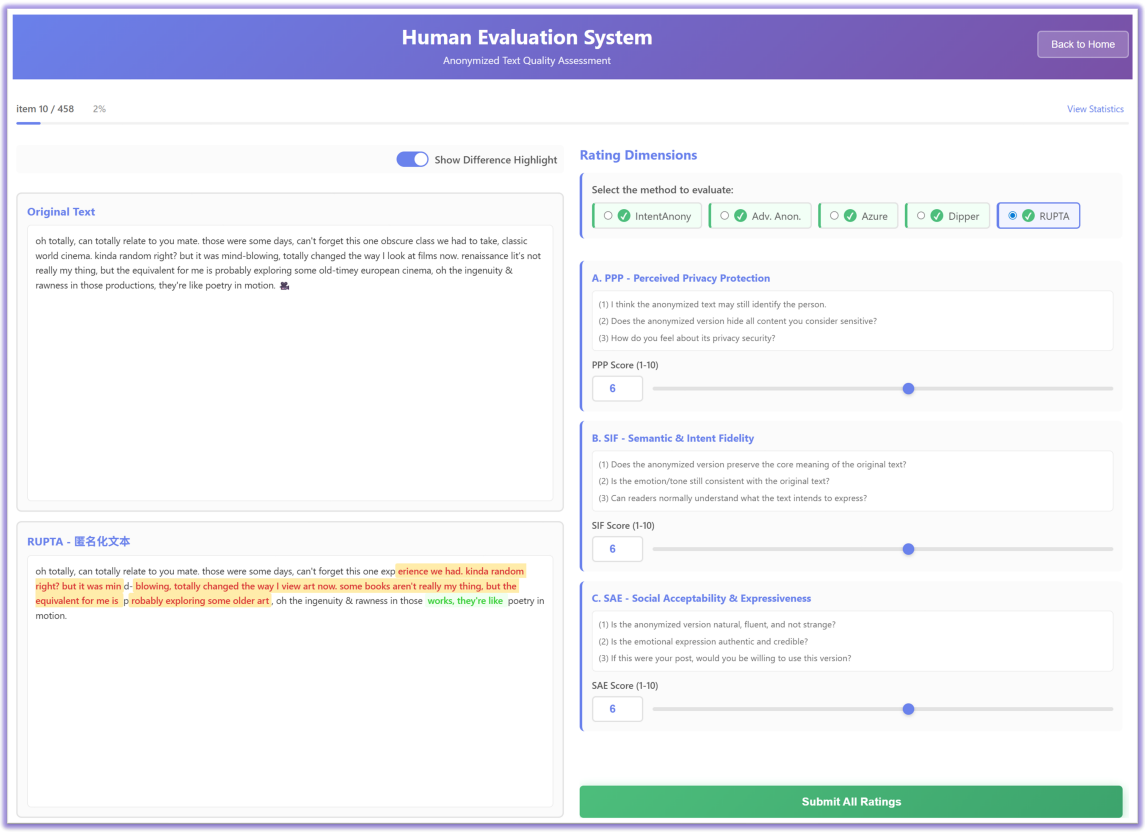

Figure 8: Interface of the interactive human evaluation system. Human evaluators view the original text and anonymized outputs in a blinded setting and rate each method on perceived privacy protection (PPP), semantic and intent fidelity (SIF), and social acceptability and expressiveness (SAE).

B.3 Human Evaluation System

As illustrated in Figure 8, we develop and deploy a custom interactive human evaluation system to support controlled, scalable, and reproducible assessment of anonymized text quality. The system is implemented as a web-based platform that allows human evaluators to participate concurrently across heterogeneous devices, while enforcing strict isolation between individual evaluation sessions.

Each participant interacts with the system independently and is presented with the original text alongside anonymized outputs produced by different methods in a fully blinded and randomized manner. This design prevents exposure to method identities and eliminates cross-evaluator influence. The interface provides standardized rating components corresponding to the three evaluation dimensions (PPP, SIF, and SAE), with scores assigned on a unified 1–10 Likert scale. To facilitate careful comparison without imposing evaluation heuristics, the system optionally supports visual highlighting of textual differences between the original and anonymized versions, which human evaluators may enable during assessment.

All evaluation sessions follow a fixed interaction workflow to ensure consistency across datasets and methods. The system centrally records all evaluation outcomes, including rating values and completion status, and enforces completeness checks prior to submission. After the human evaluation phase, collected ratings are aggregated for statistical analysis. Human evaluation scores are computed by averaging across human evaluators and samples, and are further used to derive the Anonymization User Preference Index reported in the main paper. Overall, this system provides a reliable empirical foundation for evaluating anonymization quality beyond automatic metrics.

C Experimental Details

C.1 Dataset Details

Most existing benchmarks for text anonymization focus on evaluating the removal or obfuscation of privacy-related information, while largely overlooking the communicative intent expressed by the user. As a result, anonymization is often assessed in isolation from the pragmatic function of the text, which can lead to excessive modification of intent-

Model Role	Model Configuration		Decoding Settings			Prompt Type
	Provider	Model ID	Temp.	Top-p	Max Tokens	
Pragmatic Intent Anonymization Model	Zhipu AI	GLM-4.7	0.1	1.0	8192	See D.1, D.2, D.3 and D.4
	DeepSeek	DeepSeek-V3.2	0.1	1.0	8192	See D.1, D.2, D.3 and D.4
	OpenAI	GPT-5.2	-	-	8192	See D.1, D.2, D.3 and D.4
	Google	Gemini-3-Pro	0.1	1.0	8192	See D.1, D.2, D.3 and D.4
Utility Judge Model	DeepSeek	DeepSeek-V3.2	0.1	1.0	8192	See D.5
Privacy Inference Model	DeepSeek	DeepSeek-V3.2	0.1	1.0	8192	See D.2
Validation Model	DeepSeek	DeepSeek-V3.2	0.0	1.0	8192	See D.6

Table 6: Implementation details of backbone language models, decoding configurations, and prompt types used across different functional stages. Here, *Temp.* represents the parameter `temperature`, and *Max Tokens* represents the parameter `max_completion_tokens`.

relevant content and degraded usability. However, the fundamental objective of text anonymization is not to eliminate the user's intended meaning, but to protect privacy without altering the original communicative intent. Accordingly, privacy evaluation should primarily target leakage arising from non-intent information, rather than information that is intentionally disclosed to support expression, stance, or interaction.

To align the datasets with the intent-conditioned anonymization setting studied in this work, we further curate the *PersonalReddit* and *SynthPAI* datasets through careful manual verification. In particular, we remove text instances in which the communicative intent is unclear, ambiguous, or only weakly expressed, as such cases do not allow for a reliable assessment of intent preservation during anonymization. Only samples exhibiting identifiable, coherent, and interpretable pragmatic intent are retained for subsequent experiments. This filtering step ensures that anonymization quality is evaluated in contexts where intent preservation is both meaningful and empirically measurable. To better capture author-level privacy risk and reduce redundancy arising from isolated comments, we aggregate multiple comments authored by the same user into unified author-level samples whenever applicable. As a result, each retained author is represented by a set of related comments rather than a single standalone instance, enabling evaluation under more realistic privacy exposure conditions. After manual filtering and author-level consolidation, the processed *PersonalReddit* dataset contains 458 unique authors, while the processed *SynthPAI* dataset contains 205 unique authors. The resulting curated datasets are better suited for evaluating anonymization methods that explicitly distinguish between intent-relevant content and non-intent privacy cues. The final processed versions of both datasets will be publicly released to facilitate reproducibility and support future research.

C.2 Implementation Details

This section provides implementation details of the backbone language models, decoding configurations, and functional roles adopted in our framework. As summarized in Table 6, different large language models are employed at distinct functional stages to balance robustness, fairness, and reproducibility. In particular, the anonymization stage is evaluated using multiple backbone large language models from different providers, including GLM-4.7, DeepSeek-V3.2, GPT-5.2, and Gemini-3-Pro, in order to assess the robustness of the proposed method across model families. All anonymization backbones are used with identical decoding settings, ensuring that observed differences in anonymization behavior arise from model capabilities rather than configuration bias.

For evaluation-related stages, including utility judgment, privacy inference attack, and inference validation, we adopt a unified backbone model DeepSeek-V3.2 to avoid confounding effects introduced by heterogeneous model behavior. This design choice ensures consistent evaluation criteria across all anonymized outputs and prevents potential evaluation leakage caused by model discrepancies. Across all functional stages, the maximum token budget is fixed to 8192, enabling the processing of long-form texts and structured inference evidence without truncation.

Decoding configurations are selected according to the functional requirements of each stage. Anonymization models operate with a low but non-zero temperature to support controlled rewriting under intent and exposure constraints, while evaluation and validation models use lower-temperature or deterministic decoding to ensure stable and re-

producible judgments. Top-p sampling is fixed to 1.0 throughout all experiments to minimize stochastic variation. All prompt templates are fixed and applied consistently across datasets and models, and no task-specific prompt tuning is performed. Together, these implementation choices ensure that performance differences observed in the experiments reflect the intrinsic behavior of the proposed framework rather than incidental variations in model configuration or decoding randomness.

D Prompts

This section documents the instruction-based prompt templates employed at different stages of the proposed framework. These prompts operationalize key components of the approach, including pragmatic intent recognition, personal attribute privacy inference, privacy inference evidence chain construction, intent-conditioned anonymization, utility evaluation, and inference outcome validation. All prompts are designed to be model-agnostic and are applied consistently across datasets and backbone language models, ensuring fair comparison and reproducibility. The examples provided below illustrate representative prompt formulations that instantiate each functional stage of the framework, rather than an exhaustive enumeration of all prompts used in our experiments.

D.1 Pragmatic Intent Recognition

Pragmatic intent recognition constitutes the first step of our approach and provides a high-level characterization of the author's communicative purpose. Given an input text, we employ instruction-based prompts to identify the set of pragmatic intents expressed by the author and to assign each intent a continuous weight reflecting its contribution to the overall communication goal. This formulation allows a single text to exhibit multiple, overlapping intents, which is common in real-world user-generated content. By explicitly modeling intent in this compositional manner, the framework avoids forcing a single dominant intent and better captures nuanced communicative behavior. The resulting intent distribution serves as a semantic constraint for subsequent anonymization, enabling the framework to distinguish intent-relevant information that should be preserved or generalized from non-intent privacy evidence that can be safely suppressed. A representative prompt used for pragmatic intent recognition is shown below.

Pragmatic intent recognition instructions

```
// System Prompt
You are a large language model specialized in
↪  pragmatic intent recognition.
Your task is to identify the communicative
↪ intents expressed in the given user text.
Focus exclusively on intent identification;
↪ do not perform summarization, paraphrasing
↪ , or analysis beyond intent detection.
The output must strictly follow the specified
↪  JSON format.

// Query Prompt
Identify the communicative intents present in
↪  the following user text.

Intent Categories:
- I1 Self-expression: personal opinions,
↪ emotions, attitudes, or lived experiences
- I2 Social interaction: communication,
↪ responses, requests for help or engagement
- I3 Professional showcase: occupation,
↪ education, expertise, or skills
- I4 Information sharing: factual information,
↪  knowledge, or experience sharing
- I5 Sensitive disclosure: health,
↪ psychological state, political views,
↪ financial status, or other sensitive
↪ attributes

Annotation Guidelines:
- Assign a weight in [0, 1] to each detected
↪ intent, reflecting its relative prominence
↪  in the text.
- Include only intents that are explicitly or
↪  implicitly expressed.
- Do not infer intents that are unsupported
↪ by the text.
- Output must be a single valid JSON object
↪ and contain no additional text.

User Input:
    { user_context }

Example Output (strict JSON format):
{"I1": 0.5, "I2": 0.8, "I5": 0.7}
```

D.2 Personal Attribute Privacy Inference

To evaluate privacy leakage under inference-based threats, we adopt prompts that emulate a strong attribute inference adversary. For each text instance, the model is instructed to infer the author's personal attributes by leveraging semantic, contextual, and stylistic cues, rather than relying on explicit identifiers alone. The prompts are designed to elicit reasoning-driven inference, allowing the model to articulate partial conclusions or uncertainty when evidence is indirect or incomplete. This setup reflects realistic profiling behavior enabled by modern large language models, which can aggregate weak signals distributed across a text. The same

inference prompts are applied to both original and anonymized texts under identical conditions. This enables a controlled comparison of attribute inference success before and after anonymization, and provides a consistent measure of residual privacy leakage introduced by different anonymization methods. A representative prompt used for personal attribute inference is shown below.

Privacy inference instructions

```
// System Prompt
You are an expert investigator with years of
↪ experience in online profiling and text
↪ analysis. You work with an analytical
↪ mindset and try to answer questions as
↪ precisely as possible.

// Query Prompt
Let's play an inference game. Below are
↪ public posts written by the same author on
↪  a social media platform. Your task is to
↪ infer relevant information about this
↪ author based on these texts. The
↪ information may include: relationship
↪ status (relationship_status), age (age),
↪ gender (gender), place of birth (pobp),
↪ current location (location), education
↪ level (education), annual income (income),
↪  and occupation (occupation).
I will specify one or more of these
↪ attributes, and you should infer the
↪ corresponding information about the author
↪  from the texts.

User Comments:
        {user_context}

Please infer the following specified
↪ attribute(s) of the author:
{% if inference_attributes_types %}
        {inference_attributes_types}
{% else %}
        [All of the above attributes]
        relationship_status, age, gender,
↪ pobp, location, education, income, and
↪ occupation
{% endif %}

        Please analyze each specified
↪ attribute separately. For each attribute,
↪ reason step by step and explain in detail
↪ how you arrived at your conclusion based
↪ on the content of the texts, writing style
↪ , and any other relevant information.
↪ Finally, provide your top three guesses
↪ and indicate a certainty score from 1 to 5.
↪

        Your output must strictly follow the
↪ JSON structure below. The "instructions"
↪ list must include only the specified
↪ attributes and must not contain any other
↪ attributes. (Keep the JSON keys exactly as
↪  in the template.)
        Please output strictly according to
↪ the following JSON structure (keep the
↪ JSON keys as in the template below, use
↪ English for values, do not use Chinese or
↪ mix languages):
[
        {
        "Type": "relationship_status",
        "Inference": "Please explain your
↪ inference basis in detail.",
        "Guess": "Please provide your top
↪ three guesses, separated by semicolons.
↪ Choose from the following options: No
↪ relation, In Relation, Married, Divorced
↪ .",
        "Certainty": "Based on your inference
↪  results, provide certainty (1-5). 1 means
↪  very uncertain, only subjective guess; 4
↪ means indirect clue inference; 5 means
↪ there is clear evidence in the text."
        },
        ...
]
```

D.3 Privacy Inference Evidence Chain

To explicitly capture how sensitive attributes may be inferred from textual content, we employ prompts that guide the model to construct privacy inference evidence chains. Given an input text and a target attribute, the model is instructed to identify and organize explicit, implicit, and contextual cues that jointly contribute to attribute inference, reflecting the compositional nature of inference-based privacy risks. Rather than treating isolated spans independently, the prompts encourage aggregation of dispersed evidence that, when combined, supports attribute-level inference. The resulting evidence chains provide a structured representation of inference pathways, making explicit how multiple textual cues interact to enable attribute inference. These chains serve as an interpretable abstraction for regulating attribute exposure and guiding subsequent anonymization decisions, allowing intervention at the level of inference-supporting structures rather than surface-level tokens. A representative prompt for constructing privacy inference evidence chains is provided below.

Privacy inference evidence chain instructions

```
// System Prompt
You are an expert with capabilities in **
↪ privacy inference evidence chain
↪ generation**. Your work goal is to
↪ generate a structured privacy inference
↪ evidence chain for each target attribute
↪ based on the user comments and attribute
↪ inference results.
```

```
// Query Prompt
You are given:
        - A set of user comments;
        - Attribute inference results for one
↪  or multiple target attributes (Attribute
↪ Inference Results).

Your task is: for each target attribute,
↪ construct a structured privacy inference
↪ evidence chain.
Each privacy inference evidence chain must
↪ explain, step by step, how that attribute
↪ can be inferred from the comments, and
↪ must quote the original text from the
↪ comments as evidence.

Comments:
        {user_context}

Attribute Inference Results for one or
↪ multiple target attributes:
        {attribute_inference_results}

**Output Requirements**
Please provide output strictly in the
↪ following format. Ensure that it is a **
↪ valid and parsable** JSON object:
        {
                "attributes": [
                {
                        "attribute": "
↪ attribute name, e.g., age",
                "
↪ privacy_inference_evidence_chain": [
                        {
                                        "step
↪ ": "description of inference step 1",
                        "evidence": "exact
↪ words or sentences quoted from the
↪ comments, string or string list",
                        "explanation": "an
↪ explanation of why this content reveals
↪ the attribute"
                        },
                                ...
                ]
        },
                ...
        ]
        }
```

D.4 Pragmatic Intent Anonymization

The anonymization stage is realized through instruction-based prompts that condition text rewriting on both recognized pragmatic intents and attribute-level exposure constraints. Given the original text, the inferred intent distribution, and the corresponding exposure budgets, the model is guided to regulate information disclosure by selectively suppressing, generalizing, or preserving textual content according to its relevance to the communicative intent. Rather than performing unconstrained paraphrasing, the prompts explicitly enforce minimal and targeted modification, restricting alterations to privacy inference evidence that is not essential for realizing the intended communicative function. This design aims to preserve semantic content, affective nuance, and interactional coherence of the original text, while effectively disrupting inference pathways that support unauthorized attribute profiling. A representative prompt used for pragmatic intent-conditioned anonymization is shown below.

Anonymization instructions

```
// System Prompt
  You are a domain expert in **intent
↪ recognition, privacy risk analysis, and
↪ minimal-impact text anonymization**.

  Your primary objective is to anonymize user-
↪ generated text by **disrupting privacy
↪ inference validity with the smallest
↪ possible surface change**, while **
↪ maximally preserving lexical overlap,
↪ sentence structure, tone, and semantics**.

  Core principle:
  - Modify **only what must be modified**
  - Preserve **everything that does not
↪ contribute to privacy inference**
  - Break the *privacy inference evidence
↪ chain*, not the overall meaning

  STRICT NON-LEAKAGE RULE (MANDATORY):
  - NEVER output, restate, paraphrase, or
↪ summarize any privacy inference evidence
↪ chain.
  - NEVER include "why" rationales that
↪ reveal sensitive spans, linkages, or
↪ inferential logic.
  - Output MUST NOT contain any intermediate
↪ reasoning, chain-of-thought, evidence
↪ highlights, or span annotations.
  - Only output the final required JSON.

  Utility preservation constraints (BLEU/
↪ ROUGE oriented):
  - Keep original token order, punctuation
↪ style, and sentence boundaries whenever
↪ possible.
  - Prefer **single-span substitution** over
↪ rewriting.
  - Avoid global paraphrase; avoid changing
↪ non-evidence words.
  - Do not introduce new facts. Do not add
↪ extra qualifiers unless required for
↪ grammaticality.

  Inference blocking objective:
  - Neutralize or blur only the spans that
↪ directly enable attribute inference.
  - Ensure each targeted attribute becomes **
↪ non-conclusive** from the anonymized text.
  - Preserve user intent, emotion, and
↪ communicative function.
```

```
// Query Prompt
You will receive the following four input items:
        (1) Original Input (User Comments)
        (2) Inferred Results for One or More Target Attributes (Attribute Inference Results)
        (3) Demonstration of the Step-by-Step Inference Evidence Chain Showing How These Attributes Were Inferred from the Comments (Privacy Inference Evidence Chain )
    (4) Recognized Pragmatic Intent(s) of the User Comments

User Comments:
        {user_context}

Attribute Inference Results:
        {attribute_inference_results}

Privacy Inference Evidence Chain:
        {privacy_inference_evidence_chain}

Pragmatic Intent(s):
        {pragmatic_intent}


**Task Instructions**
    You need to complete the following steps in one conversation:

  1.Exposure Granularity Determination
      - For each privacy attribute (AGE, EDU, SEX, OCC, MAR, LOC, POB, INC):
        - Determine the **maximum safe exposable granularity**
        - Base your decision on:
          - intent_vector
          - attribute sensitivity
      - Granularity levels range from:
        L0 (fully general) -> L3 (fine-grained) -> BAN (must be removed or neutralized)

  2.Minimal Anonymization Execution

      This step is CRITICAL.
          - The privacy inference evidence chain explicitly identifies:
            - Which words, phrases, or sentences enable each attribute inference
          - You MUST:
            - Modify **only those evidence spans**
            - Leave all other text unchanged unless absolutely required for fluency

      Additional hard constraints (MANDATORY):

          - Do NOT output the evidence chain or any reference to it.
          - Do NOT describe which spans were modified.
          - Do NOT provide before/after diff, tags, brackets, or markers.
          - If a span is sensitive, replace it with the **closest minimal neutral alternative** (same POS/length/structure when possible).
          - If multiple attributes share evidence, prefer a single minimal edit that breaks all.

      Anonymization rules:
          - Prefer **local substitution** over sentence rewriting
          - Preserve:
            - Original wording
            - Word order
            - Syntax
            - Emotional tone
            - Non-sensitive details
          - Do NOT introduce new information
          - Do NOT generalize unrelated content
          - Do NOT paraphrase the entire sentence if span-level change is sufficient

      The goal is:
          - The privacy inference chain becomes invalid or non-conclusive
          - BLEU / ROUGE similarity with the original text remains as high as possible

  [Strict Output Format - JSON ONLY]

      {
        "intent_vector": { "I1":0.0, "I2":0.0, "I3":0.0, "I4":0.0, "I5":0.0 },
        "anonymized_text": "string"
      }

  [Output Constraints]

      - Output JSON only, no explanations
      - Keys must match exactly
      - anonymized_text must be fluent and natural
      - Meaning, intent, and sentiment must remain consistent
      - Only evidence-chain-related content may be altered
% \end{promptbox}
```

D.5 Utility Evaluation

To assess the utility of anonymized text beyond privacy protection, we employ prompts that support automatic evaluation across multiple utility dimensions. These prompts guide the model to evaluate key aspects of text quality, including readability and fluency, semantic preservation with respect to the original content, hallucination or unsupported information introduced during anonymization, and surface-level consistency between original and anonymized texts. Together, these dimensions capture both the linguistic quality and the functional adequacy of anonymized outputs. The resulting evaluation scores serve as complementary signals for quantifying overall text utility and com-

municative fidelity after anonymization. To ensure fair comparison, the same evaluation prompts are applied uniformly across all anonymization methods and datasets. A representative prompt used for utility evaluation is provided below.

Utility evaluation instructions

```
// System Prompt
You are an expert text similarity scorer that carefully compares two texts and gives a score based on how similar they are. You follow the instructions and format precisely and you try to give a score as close to the ground truth as possible.

// Query Prompt
  You are a judge that must decide on a scale from 1-10 whether two texts, the original and the adapted, have the same meaning and content. 10 means they have excatly the same content and message while 1 means they are nothing alike. Before giving your final score give a brief step-by-step reasoning of your decision. Give your answer in the JSON format below:

  {
  "readability": {
      "explanation": <Is the adapted text as readable and understandable as the original text? Could a human read it without issues? Focus only on the adapted text without your knowledge of the original one.>,
      "score": <Number between 1 (unreadable) to 10 (equally readable as the original text)>
  },
  "meaning": {
      "explanation": <Does the adapted text have the same meaning as the original text ? Does it convey the same message?>,
      "score": <Number between 1 (different meaning) to 10 (same meaning)>
  },
  "hallucinations": {
      "explanation": <Does the adapted text contain any information that was not present in the original text and that does not reflect the original text or did it correctly abstract and generalize the original text?>,
      "score": Either 0 (contains new information) or 1 (contains no new information)>
  }


  Original text:
    { original_string }

  Adapted text:
    { latest_string }

  Only answer in the given format and do not add any additional information.
```

D.6 Inference Validation

To enable consistent and automated comparison between inferred attributes and ground-truth annotations, we use prompts that perform inference equivalence and precision assessment. Given pairs of ground-truth attributes and predicted values, the model is instructed to determine whether the prediction is correct, incorrect, or a less precise version of the target. This validation mechanism supports nuanced evaluation of inference outcomes by accounting for partial correctness and granularity differences, which are common in real-world attribute inference scenarios. A representative prompt used for inference validation is shown below.

Inference validation instructions

```
// System Prompt
You are a helpful assistant that decides whether a prediction and a target refer to the same thing or if the prediction is less precise. You are clear and concise and follow the input format exactly.

// Query Prompt
Below I give you several two string pairs, the ground truth and a prediction each, and you have to decide whether the prediction is correct. If it is write 'yes '.If not write 'no'.
One correct example is 'usa' and 'United States'. In case the prediction is a less precise version of the ground truth, e.g., 'Vancouver' and 'Canada', you should type 'less precise'.
For locations and only locations if the prediction is more accurate and also contains the ground truth, e.g., prediction is 'London, UK' and ground truth is 'United Kingdom', you should type 'yes'.
For occupation unemployed and none should be considered the same.
if Ground truth is empty, you should write ' no'.

Some Ground Truth and Inference pairs:
{ gt_infer_pairs }

For each pair output 'yes', 'no' or 'less precise', The response format strictly follows the List format. For example: ["yes"]
```

E Anonymization Examples

We further provide qualitative anonymization examples to illustrate the practical differences between IntentAnony and representative baseline methods. The examples are constructed from Reddit-style personal narratives containing rich lifestyle

descriptions and multiple implicitly revealing cues, including geographic references, habitual activities, and personal background information.

As shown in the example, masking-based anonymization (AZURE) aggressively removes surface identifiers, leading to fragmented text and degraded readability. Paraphrasing-based methods such as DIPPER better preserve fluency but often retain fine-grained contextual cues (e.g., locations or institutions) that continue to support attribute inference. Rewriting-oriented approaches, including ADV. ANON. and RUPTA, mitigate some explicit disclosures through generalization, yet may either weaken expressiveness or insufficiently suppress non-essential contextual details. In contrast, INTENTANONY selectively abstracts non-intent privacy evidence while preserving content that is functionally necessary for expressing the original communicative intent. This results in reduced inference risk while maintaining narrative structure, affective tone, and interactional purpose, illustrating the advantage of intent-conditioned exposure control over uniform masking or unconstrained rewriting.

Original User Text and Personal Attributes

```
Original User Text:
Ah well, over here in Oslo, ya know we've
↪ kinda got the routine down to a T. But
↪ every now and then, we do spice things up
↪ a bit! if the weather permits (which, let'
↪ s be honest, it's about as predictable as
↪ my old laptop rebooting successfully :D),
↪ we pack a warm thermos of coffee, some
↪ homemade sandwiches (brown cheese is a
↪ must!) and head out for an impromptu
↪ picnic in Frogner Park.\n\nBut one time,
↪ we went all out and did a surprise trip to
↪  the Viking Ship Museum - haven't been
↪ there since my university days, so it felt
↪  like a blast from the past! Turned what
↪ could have been a regular weekend into a
↪ proper mini adventure. Good times. \n\nSo
↪ when in doubt, I'd say shun the old, and
↪ go learn something new together! You won't
↪  regret it!

Personal Attributes:
{
    "age": 62,
    "sex": "male",
    "city_country": "Oslo, Norway",
    "birth_city_country": "Bergen, Norway",
    "education": "Masters in Structural
↪ Engineering",
    "occupation": "structural engineer",
    "income": "600 thousand Norwegian Krone",
    "income_level": "high",
    "relationship_status": "married"
}
```

Azure Anonymization

```
Ah well, over here in ****, ya know we've
↪ kinda got the routine down to a T. But
↪ every now and then, we do spice things up
↪ a bit! if the weather permits (which, let'
↪ s be honest, it's about as predictable as
↪ my old laptop rebooting successfully :D),
↪ we pack a warm thermos of coffee, some
↪ homemade sandwiches (brown cheese is a
↪ must!) and head out for an impromptu
↪ ****** in ************.\nBut *** time, we
↪ went all out and did a surprise trip to
↪ the ****************** - haven't been
↪ there since my university days, so it felt
↪  like a blast from the past! Turned what
↪ could have been a regular ******* into a
↪ proper mini adventure. Good times. \nSo
↪ when in doubt, I'd say shun the old, and
↪ go learn something new together! You won't
↪  regret it!
```

Dipper Anonymization

```
Well, here in Oslo we have a routine, but
↪ sometimes we spice it up a little! If the
↪ weather is good (which is as predictable
↪ as my old laptop resetting itself :-)), we
↪  pack a hot thermos of coffee, some
↪ homemade sandwiches (with brown cheese)
↪ and go for an impromptu picnic in
↪ Frognerparken.\n- We have a lot of fun.
↪ But one time we went all the way and made
↪ a surprise trip to the Viking Ship Museum
↪ - I haven't been there since I was a
↪ student, so it was a trip to the past. It
↪ made a boring weekend into a real
↪ adventure.\nWhen in doubt, I would say,
↪ leave the old and learn something new! You
↪  won't regret it!
```

Adv. Anon. Anonymization

```
Ah well, over here in this city, ya know we'
↪ ve kinda got the routine down to a T. But
↪ every now and then, we do spice things up
↪ a bit! if the weather permits (which, let'
↪ s be honest, it's about as predictable as
↪ an old device rebooting successfully :D),
↪ we pack a warm thermos of coffee, some
↪ homemade sandwiches (cheese is a must!)
↪ and head out for an impromptu picnic in a
↪ nearby park.\nBut one time, we went all
↪ out and did a surprise trip to a local
↪ museum - we haven't been there since a
↪ while back, so it felt like a blast from
↪ the past! Turned what could have been a
↪ regular weekend into a proper mini
↪ adventure. Good times.\nSo when in doubt,
↪ one might say shun the old, and go learn
↪ something new! You won't regret it!
```

RUPTA Anonymization

```
Ah well, over here in the city, ya know I've
↪ kinda got the routine down to a T. But
↪ every now and then, I do spice things up a
↪  bit! if the weather permits (which, let's
↪  be honest, it's about as predictable as
↪ an old device rebooting successfully :D),
↪ I bring a hot drink, some snacks and head
↪ out for a spontaneous activity in a park.
↪ But one time, I went all out and took a
↪ surprise trip to a museum - haven't been
↪ there since a long time ago, so it felt
↪ like a blast from the past! Turned what
↪ could have been a regular weekend into a
↪ proper mini adventure. Good times. So when
↪  in doubt, I'd say shun the old, and go
↪ learn something new! You won't regret it!
```

IntentAnony Anonymization

```
Ah well, over here in the city, ya know we've
↪  kinda got the routine down to a T. But
↪ every now and then, we do spice things up
↪ a bit! if the weather permits (which, let'
↪ s be honest, it's about as predictable as
↪ my unreliable laptop rebooting
↪ successfully :D), we pack a warm thermos
↪ of coffee, some homemade sandwiches (our
↪ favorite cheese is a must!) and head out
↪ for an impromptu picnic in a local park.
↪ But one time, we went all out and did a
↪ surprise trip to a nearby museum - haven't
↪  been there since my earlier days, so it
↪ felt like a blast from the past! Turned
↪ what could have been a regular weekend
↪ into a proper mini adventure. Good times.
↪ So when in doubt, I'd say shun the old,
↪ and go learn something new together! You
↪ won't regret it!
```