

Privacy-R1: Privacy-Aware Multi-LLM Agent Collaboration via Reinforcement Learning

Zheng Hui[✉], Yijiang River Dong[✉], Sanhanat Sivapiromrat[✉]

Ehsan Shareghi[✉], Nigel Collier[✉]

[✉] University of Cambridge, [✉] University College London
{zh2483}@columbia.edu,
{yd358, ss3229, es776, nhc30}@cam.ac.uk

Abstract

When users submit queries to Large Language Models (LLMs), their prompts can often contain sensitive data, forcing a difficult choice: Send the query to a powerful proprietary LLM providers to achieving state-of-the-art performance and risk data exposure, or relying on smaller, local models guarantees data privacy but often results in a degradation of task performance. Prior approaches have relied on static pipelines that use LLM rewriting, which shatters linguistic coherence and indiscriminately removes privacy-sensitive information, including task-critical content. We reformulate this challenge (Privacy-Conscious Delegation) as a sequential decision-making problem and introduce a novel reinforcement learning (RL) framework called Privacy-R1 to solve it. Our framework trains an agent to dynamically route text chunks, learning a policy that optimally balances the trade-off between privacy leakage and task performance. It implicitly distinguishes between replaceable Personally Identifiable Information (PII) (which it shields locally) and task-critical PII (which it strategically sends to the remote model for maximal utility). To validate our approach in complex scenarios, we also introduce a new medical dataset with high PII density. Our framework achieves a new state-of-the-art on the privacy-utility frontier, demonstrating the necessity of learned, adaptive policies for deploying LLMs in sensitive environments. Dataset can be found at: <https://github.com/zackhuiiii/Privacy-R1>.

1 Introduction

As users increasingly rely on Large Language Models (LLMs) (Brown et al., 2020) across a wide range of domains, their applications now extend far beyond basic text generation. These include synthetic data generation (Hui et al., 2024, 2025b), professional and creative writing (Chakrabarty et al., 2024), educational support and tutoring (Razafinirina et al., 2024), legal drafting (Kolt et al., 2026),

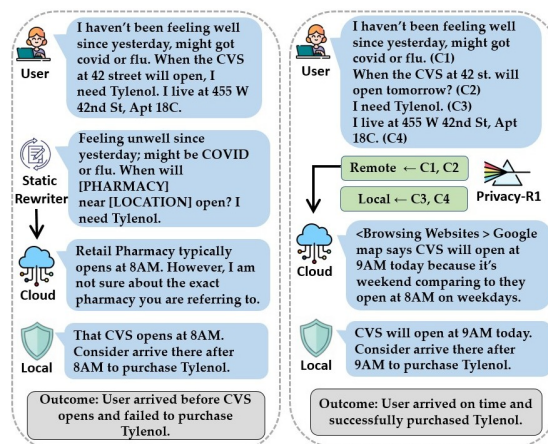


Figure 1: A static rewriter (left) fails by redacting task-critical information. Our Privacy-R1 framework (right) succeeds by learning a policy to route sensitive and non-sensitive query chunks to the appropriate model, preserving both privacy and task performance.

personal assistance systems (Li et al., 2024; Dong et al., 2026a) and memory-augmented systems (Huang et al., 2026; Hui et al., 2026; Zhang et al., 2026) that adapt to user preferences over time. As a result, user prompts often contain sensitive information, ranging from personal identifiers and confidential business data to health and financial details (Ai et al., 2024). This presents a dilemma: sending a query to a powerful, state-of-the-art remote API risks the exposure private data, while processing it with a secure, locally-hosted smaller model that may yields a lower-quality response. This conflict between utility and privacy is a primary obstacle to deploying LLMs in critical domains like healthcare and finance (Rathod et al., 2025; Hui et al., 2025a), making the development of trustworthy solutions a crucial area of research. To formally address this trade-off, recent work has proposed the task of Privacy-Conscious Delegation (Li et al., 2025). The goal is to create a system where a secure local model acts as an intelligent proxy, leveraging the power of a remote API to fulfill a user's request without leaking their Personally Identifiable

Information (PII). Initial approaches to this task have employed a static, monolithic prompt rewriting pipeline. This paradigm attempts to paraphrase the entire user query to remove all PII before sending it to the remote API. However, this strategy is inherently brittle. By treating the entire query as a single unit and removing all PII, it fails to make nuanced decisions, leading to two catastrophic failure modes. The first is discourse coherence failure, where removing entities severs critical linguistic links within the text. The second is utility collapse as illustrated in Figure 1, where the PII is integral to the user’s goal, and its removal renders the query unanswerable. The core limitation of these approaches is that they are based on a fixed, context-blind policy. We argue that effective delegation is not a one-shot text manipulation task, but rather a sequential decision-making problem where the optimal action depends on both the immediate context and the overall user objective. To solve this, we introduce a novel reinforcement learning (RL) framework that explicitly learns to balance the competing pressures of privacy and performance. Our framework trains a stateful policy agent to make granular, chunk-by-chunk routing decisions. By optimizing for a reward function that combines downstream task performance with a penalty for privacy leakage, our agent learns a dynamic and pragmatic policy. It learns when to shield information and when a calculated privacy risk is necessary to achieve a high-quality outcome for the user. Our primary contributions are threefold:

- We introduce a novel RL framework call Privacy-R1 featuring a lightweight stateful policy agent that dynamically route content between local and remote models. By optimizing a composite reward, it implicitly learns to differentiate between replaceable PII (which it shields) and task-critical PII (which it may strategically reveal for a significant utility gain).
- We contribute a new, challenging medical PII dataset, featuring a higher density of PII entities. This benchmark is designed to stress-test the contextual reasoning capabilities of privacy-preserving systems.
- Our experiments show that Privacy-R1 outperforms baselines, charting a new state-of-the-art on the privacy-utility frontier. The performance gains are especially pronounced on our more complex dataset, underscoring the necessity of adaptive, learned frameworks for building the

next generation of trustworthy AI systems.

2 Related Works

Training-Time vs. Inference-Time Privacy While much privacy research focuses on training-time risks like data memorization and federated learning (Kandpal et al., 2022; Carlini et al., 2022; Wang et al., 2024; Ruzzetti et al., 2025; Dong et al., 2026b), protecting user data at inference time is a critical challenge (Miresghallah et al., 2024). Classical approaches (Herwanto et al., 2021; Mahendran et al., 2021; Pamarthi, 2024) often rely on Named Entity Recognition (NER) models to perform static redaction method that frequently degrades semantic coherence and downstream task utility. The task of Privacy-Conscious Delegation (Li et al., 2025) was proposed to use LLM to paraphrase the entire user query to remove PII. **Our work accepts this task definition but argues that a static, global rewrite is suboptimal. We instead propose a dynamic, routing approach.**

Reinforcement Learning for Privacy-Aware Control Research at the intersection of RL and privacy has largely concentrated on two main directions. The first focuses train an RL to protect sensitive information against inference or reconstruction attacks (Andreoletti et al., 2020; Qiao and Wang, 2023; Mo et al., 2024; Tan et al., 2025). The second line of work applies RL to generative models, where rewards are formulated to balance the quality or utility of generated outputs (e.g., text, images) with privacy metrics (Shaik et al., 2024; Saha, 2025; Hore et al., 2025). In contrast, we are not primarily defending against an attacker or modifying the text itself; instead, we train an agent to learn a **meta-policy for information routing**. To our knowledge, this is the first work to use RL to learn a stateful, dynamic delegation policy that optimally balances the privacy-utility trade-off at a sub-prompt level.

3 Problem formulation

We address the task of **Privacy-Conscious Delegation** (Li et al., 2025). The task is designed to navigate the trade-off between the performance of powerful, remote LLMs and the security of local, trusted models. The setup involves two models:

- M_{LOCAL} : A trusted, secure LLM model with potentially weaker capabilities, which can be run locally or on a private server.

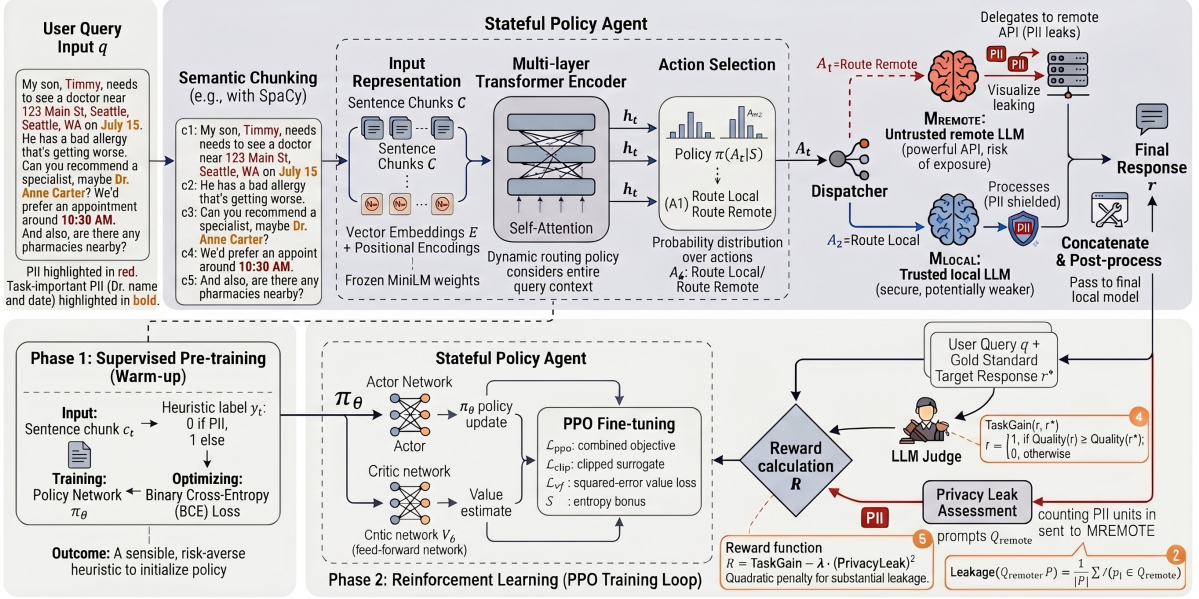


Figure 2: Privacy-R1 Framework.

- M_{REMOTE} : An untrusted, powerful LLM model accessed via Proprietary API.

The system is given a user query q , which may contain a set of one or more Personally Identifiable Information (PII) units, denoted as $P = \{p_1, p_2, \dots, p_k\}$. The fundamental goal is to generate a final response r that is of high quality, while ensuring that the information delegated to M_{REMOTE} contains as little PII as possible. To quantify success in this task, we adopt the evaluation metrics proposed in the original work:

Quality Preservation The performance of a system is measured by its ability to produce a response r that is comparable in quality to a gold-standard target response, r^* . This target is defined as the output from the powerful M_{REMOTE} when given the original, unaltered query q . The quality is determined by a scoring function, J_{qual} :

$$\text{Quality}(r) = J_{\text{qual}}(r, r^*) \quad (1)$$

This function is typically implemented using an LLM-as-a-judge to assess the semantic equivalence and utility of the generated response relative to the target.

Privacy Leakage The privacy cost is measured as the fraction of PII units from the original query that are exposed to the untrusted remote model. Let Q'_{remote} be the complete set of all prompts synthesized and sent to M_{REMOTE} during the delegation

process. The leakage is then formally defined as:

$$\text{Leakage}(Q'_{\text{remote}}, P) = \frac{1}{|P|} \sum_{i=1}^k \mathbb{I}(p_i \in Q'_{\text{remote}}) \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function, which evaluates to 1 if the PII unit p_i is present in any prompt sent to the remote model, and 0 otherwise.

The central challenge of Privacy-Conscious Delegation is thus to design a system that can effectively navigate the tension between these two competing objectives, maximizing the quality score while minimizing the leakage score.

4 Privacy-R1

As shown in Figure 2, to solve the problem defined in Section 3, we reframe Privacy-Conscious Delegation as a sequential decision-making problem. We introduce a RL framework called **Privacy-R1** short for **PRIVACY** preservation policy agent for dynamic delegation via **Reinforcement Learning** that trains a policy agent to learn an optimal, dynamic routing policy. This section details the architecture of our agent, the design of our reward function, and the training procedure used to optimize the policy.

4.1 Framework Overview

Our framework processes a user query q in three main stages. First, we perform semantic chunking to segment the query into a sequence of meaningful units. Second, our Stateful Policy Agent processes this sequence to make a routing decision for each

chunk. Finally, the chunks are dispatched, and their outputs are combined to form the final response.

Semantic Chunking The first step is to segment the user query q into a sequence of coherent units for the policy agent. We employ a semantic chunking strategy. Specifically, we use a sentence segmentation model from SpaCy to split the query q into a sequence of sentences $C = (c_1, \dots, c_n)$. This ensures that each input to our agent is a linguistically complete thought, which provides a much cleaner learning signal and preserves the contextual relationships necessary for learning a robust routing policy.

Dynamic Routing and Response Generation

The Policy Agent processes the entire sequence of chunks C to assign an action A_t (route local or remote) to each chunk c_t . The chunks are then dispatched to the corresponding models (M_{LOCAL} or M_{REMOTE}), and their individual outputs are concatenated and pass to local model in order to form the final response r .

4.2 The Policy Agent

To effectively model the complex, long-range dependencies within a user query, our policy agent is implemented using a lightweight **Transformer** architecture. This design allows the agent to consider the entire query context when making a decision for any individual chunk. The parameters of this Transformer agent are the target of our training procedure.

Input Representation We use a pre-trained MiniLM model with its weights **frozen** to serve as a high-quality feature extractor. The full sequence of sentence chunks C is first converted into a sequence of dense vector embeddings $E = (e_1, \dots, e_n)$. To provide positional context, we add standard sinusoidal positional encodings to these embeddings.

Contextualization with Self-Attention The sequence of embeddings E is then passed through a multi-layer Transformer encoder. The self-attention mechanism computes a new representation, h_t , for each chunk c_t that is contextualized by all other chunks in the sequence:

$$(h_1, \dots, h_n) = \text{Transformer}(e_1, \dots, e_n) \quad (3)$$

This global context is critical for making informed decisions, such as understanding that a pronoun in the final chunk refers to a PII entity in the first.

Action Selection For each contextualized output vector h_t , a shared linear layer followed by a softmax activation produces the probability distribution over the two possible actions. This distribution is our policy π for each step:

$$\pi(A_t|S) = \text{Softmax}(W \cdot h_t + b) \quad (4)$$

where the state S implicitly represents the entire sequence of chunks. During training, an action A_t for each chunk is sampled from this distribution to encourage exploration.

4.3 Reward Design: Balancing Privacy and Performance

The reward signal is calculated after a full sequence of actions has been taken. The reward function R explicitly models the privacy-utility trade-off.

TaskGain This term corresponds to the $\text{Quality}(r)$ metric and serves as the primary performance signal for our agent. Following the established protocol from Papillon (Li et al., 2025), we treat TaskGain as a **binary signal**. An LLM-as-a-judge determines if the system’s final response r is of ‘equivalent or better quality’ than the target response r^* . The reliability of this specific automated evaluation judge for this task has been previously validated with human studies in Papillon, which found substantial agreement between the LLM judge and human annotators. By adopting this validated judge, we ensure a reliable reward signal for our agent. The reward is formally defined as:

$$\text{TaskGain}(r, r^*) = \begin{cases} 1, & \text{if } \text{Quality}(r) \geq \text{Quality}(r^*) \\ 0, & \text{otherwise.} \end{cases}$$

To further mitigate known biases, we present the responses to the judge in a randomized order during the reward calculation process.

PrivacyLeak This term corresponds to the $\text{Leakage}(Q'_{\text{remote}}, P)$ metric, measuring the fraction of PII units exposed to M_{REMOTE} .

Modeling Catastrophic Risk with a Non-Linear Penalty To reflect the accelerating nature of real-world privacy risk, we introduce a **quadratic penalty** for leakage. This penalizes substantial leakage far more harshly than minor, incidental exposure. Our final reward function is:

$$R = \text{TaskGain} - \lambda \cdot (\text{PrivacyLeak})^2 \quad (5)$$

The hyperparameter λ controls the agent’s aversion to privacy risk. By default, we set $\lambda = 5.0$. We also evaluate different levels of λ in 6.4.

4.4 Policy Training

We train the policy agent in two phases: a supervised pre-training phase to provide a strong initialization, followed by a reinforcement learning fine-tuning phase to optimize for the reward.

Phase 1: Supervised Pre-training (Warm-up)

Training an RL agent from a random initialization is challenging due to the sparse and delayed nature of the reward signal. We therefore pre-train our policy network by treating it as a sequence-to-sequence binary classifier, teaching it to mimic a sensible, risk-averse heuristic. For each sentence chunk c_t in our training corpus, we generate a heuristic label y_t : if the chunk contains any PII, $y_t = 0$ (route local); otherwise, $y_t = 1$ (route remote). The policy network π_θ is then trained to predict this sequence of labels. The forward pass computes a probability $p_t = \pi_\theta(A_t = 1|S)$ for each chunk. The parameters θ of the Transformer agent are optimized by minimizing the binary cross-entropy (BCE) loss, summed over all chunks in the sequence. We use the Adam optimizer for this phase. The resulting pre-trained agent, $\pi_{\theta_{\text{st}}}$, serves as the starting point for RL fine-tuning, already equipped with a conservative policy.

Phase 2: Reinforcement Learning with PPO

We fine-tune the pre-trained policy using Proximal Policy Optimization (PPO) (Schulman et al., 2017). The policy agent serves as the **actor** (π_θ), and a separate feed-forward network serves as the **critic** (V_ϕ). The training is iterative. In each iteration, we first perform a rollout step, where the actor policy interacts with a batch of queries to collect trajectories of states, actions, and log-probabilities. For each completed trajectory, we compute the final episodic reward R and then calculate the advantage for each step, $\hat{A}_t = R - V_\phi(s_t)$, where V_ϕ is the critic’s value estimate. These advantages, which indicate the relative quality of an action, are normalized for training stability. The actor and critic are then updated by optimizing a combined objective function. This objective consists of the PPO policy loss, a value function loss for the critic, and an entropy bonus to encourage exploration. The

complete objective function to be maximized is:

$$\mathcal{L}_{\text{PPO}}(\theta, \phi) = \mathbb{E}_t [L_t^{\text{CLIP}}(\theta) - c_1 L_t^{\text{VF}}(\phi) + c_2 S[\pi_\theta](s_t)] \quad (6)$$

where $L_t^{\text{CLIP}}(\theta)$ is the PPO clipped surrogate objective that constrains the magnitude of the policy update. It is defined as:

$$L_t^{\text{CLIP}}(\theta) = \min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \quad (7)$$

Here, $r_t(\theta)$ is the probability ratio between the new and old policies. The other terms are the squared-error value function loss $L_t^{\text{VF}}(\phi)$ and the policy’s entropy S . This process allows the agent to safely and efficiently learn a sophisticated routing policy that maximizes the expected reward.

5 Experimental Setup

5.1 Datasets

PUPA To benchmark our method against prior work, we first train and evaluate on the **PUPA** dataset (Li et al., 2025). PUPA is constructed from real-world user-LLM interactions and contains naturalistic PII across a variety of general-domain topics. The dataset is divided into two primary subsets used for training and evaluation: **PUPA-New** and **PUPA-TNB**. We use the official splits and statistics as reported by the authors for our experiments.

Med-PCD While PUPA provides a valuable general-domain benchmark, specialized domains like healthcare present unique challenges, including a higher density of interconnected PII and more severe consequences for privacy breaches. To stress-test our framework in such a scenario, we introduce the **Medical Privacy-Conscious Delegation (Med-PCD)** dataset. Med-PCD is based on publicly available **MedDialog** (Zeng et al., 2020), which contains anonymized patient-doctor diagnostic conversations. To create Med-PCD, we employed GPT5 to synthetically yet realistically inject a diverse set of PII into these anonymized dialogues. We used a few-shot prompting strategy (see Appendix A for the full prompt) to guide the model in weaving entities such as patient names, doctor names, clinic locations (e.g., hospitals, departments), specific dates, and medical record numbers (MRNs) into the conversational turns. The goal was not merely to insert PII, but to create coherent and contextually rich narratives where the PII is naturally integrated.

A critical component for our RL framework is the ground-truth target response r^* , which is used to calculate the TaskGain reward. To generate these targets, we submitted each complete, PII-injected query from Med-PCD to a powerful proprietary model (GPT5). This response, generated with full access to all PII, represents the quality ceiling that our privacy-preserving system aims to match. To rigorously assess the quality and realism of our synthetic generation process, we conducted a manual validation study. Three annotators evaluated a randomly selected subset of 240 instances for contextual coherence and logical consistency. The study revealed a high success rate, with an average of 98.8% of instances being judged as pass. This validation confirms the reliability of our automated data creation pipeline. The inter-annotator agreement was substantial (Fleiss’ Kappa $\kappa = 0.89$), indicating that our quality criteria were consistently interpreted. The full details of our annotation guidelines and validation study are provided in Appendix B. The final Med-PCD dataset consists of 1020 instances, which we split into a training set of 816 instances, and test sets of 204 instances. Table 1 provides a statistical comparison between PUPA and Med-PCD, highlighting the increased complexity of our new dataset. Table 2 illustrates the transformation from an original anonymized dialogue to an injected Med-PCD instance, showcasing the complexity introduced.

Metrics	PUPA	Med-PCD (Ours)
# Instances	901	1020
Domain	General	Medical
Avg. # PII	2.90	4.6
Avg. Query Len.	1352.0	1533.2
Avg. Resp. Len.	1553.7	1920

Table 1: Statistical comparison of the datasets. The PUPA column shows weighted averages across its splits. Avg length in Char. Med-PCD is designed to be more challenging, with longer queries and a significantly higher density of PII per instance.

5.2 Models

LLM Configurations Following prior work (Li et al., 2025), we test all methods with a fixed remote model, GPT-4o-mini (Achiam et al., 2023) as M_{REMOTE} , while systematically varying the local model, M_{LOCAL} . We select a diverse suite of recent open-source models for M_{LOCAL} : **Llama-3.1-8B-Instruct** (Touvron et al., 2023), **Llama-3.2-3B-Instruct** (Grattafiori et al., 2024), **Llama-3.2-**

1B-Instruct (Grattafiori et al., 2024), **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023), and **Qwen2-7B-Instruct** (Yang et al., 2025).

Baselines We compare our proposed framework against four strong baselines for each of the local model configurations. **Always-Local** and **Always-Remote** establish the performance bounds by exclusively using one model. **PAPILLON** (Li et al., 2025), a static, monolithic prompt rewriting pipeline. Finally, to isolate the benefit of reinforcement learning, we include our **Heuristic Router (SFT-only)**, which is our policy agent after only completing the supervised pre-training phase.

5.3 Evaluation and Implementation

Metrics and Protocol We evaluate all systems on two primary metrics: **Quality Preservation (%)**, the percentage of responses judged equivalent or better than a target, and **Privacy Leakage (%)**, the average fraction of PII units exposed to the remote model.

Implementation Details Our framework is implemented in PyTorch, using SpaCy for sentence segmentation. Our policy agent is a 2-layer Transformer operating on frozen embeddings from all-MiniLM-L6-v2. The SFT (warm-up) phase is executed for 1 epochs with a batch size of 32 and a learning rate of 3×10^{-4} . Subsequently, RL with PPO is performed with a learning rate of 1×10^{-5} with batch size 64. Training is conducted for a maximum of 256 steps with a clipping parameter $\epsilon = 0.2$ and an entropy bonus coefficient of 0.01. All experiments are run on NVIDIA H200 GPUs.

6 Results and Analysis

In this section, we present the empirical evaluation of our RL framework. We first compare its performance against all baselines on both the PUPA and our new Med-PCD. Subsequently, we provide an analysis of the learning dynamics and conduct a series of ablation studies to dissect the specific contributions of our method’s key components.

6.1 Main Results

Our primary goal is to demonstrate that a learned, dynamic policy can achieve a better balance between Quality Preservation and Privacy Leakage than static approaches. The main results across a diverse suite of local models are summarized in Table 3. The findings are clear: our RL framework

Stage	Dialogue Turn (Patient’s Message)
Original (from MedDialog)	traveled 2wks ago from fl. to pa. 68 wf. has had fever of 100, chills at night and some coughing for 5ds. tested negative for flu and x rays of lungs were clear. coronavirus? scarce
After Injection (Med-PCD)	My mother, Carol , 68, returned from her trip two weak ago. She flew on UA2401 from Orlando to phil on 10/21 . For the last 5 days, she’s had a fever of 100 and night chills. Her recent flu test at the Jefferson Health clinic on Chestnut St was negative and x rays of lungs were clear. Could this be coronavirus? She does not have insurance.

Table 2: An example illustrating the creation of a Med-PCD instance. The original, anonymized text from MedDialog is transformed by synthetically injecting multiple, contextually-aware PII entities (highlighted in red) to create a realistic and challenging scenario.

Method / System	PUPA-TNB			Med-PCD		
	Qual. (%) \uparrow	Leak. (%) \downarrow	Δ (%)	Qual. (%) \uparrow	Leak. (%) \downarrow	Δ (%)
GPT-4o-mini [Unredacted]	88.2	100.0	–	91.5	100.0	–
GPT-4o-mini [Redacted]	77.2	0.0	–	72.3	0.0	–
<i>M_{LOCAL}: Llama-3.2-1B-Instruct</i>						
Always-Local	41.2	0.0		25.5	0.0	
PAPILLON	<u>58.0</u>	<u>39.3</u>		<u>45.1</u>	<u>42.5</u>	
Heuristic Router (SFT-only)	52.5	2.0		40.8	3.5	
Privacy-R1	62.5	25.0	+4.5 / +14.3	75.3	18.2	+30.2 / +24.3
<i>M_{LOCAL}: Llama-3.2-3B-Instruct</i>						
Always-Local	57.3	0.0		40.1	0.0	
PAPILLON	<u>60.9</u>	<u>24.9</u>		<u>58.5</u>	<u>28.1</u>	
Heuristic Router (SFT-only)	59.1	1.8		55.6	3.1	
Privacy-R1	65.2	19.5	+4.3 / +5.4	81.0	15.4	+22.5 / +12.7
<i>M_{LOCAL}: Llama-3.1-8B-Instruct</i>						
Always-Local	71.8	0.0		55.0	0.0	
PAPILLON	<u>85.5</u>	<u>7.5</u>		<u>82.0</u>	<u>9.2</u>	
Heuristic Router (SFT-only)	78.5	1.0		70.5	2.5	
Privacy-R1	86.1	4.5	+0.6 / +3.0	89.5	5.1	+7.5 / +4.1
<i>M_{LOCAL}: Mistral-7B-Instruct-v0.3</i>						
Always-Local	75.7	0.0		58.3	0.0	
PAPILLON	<u>77.6</u>	<u>11.9</u>		<u>74.5</u>	<u>14.0</u>	
Heuristic Router (SFT-only)	76.8	1.2		69.1	2.8	
Privacy-R1	80.2	8.1	+2.6 / +3.8	87.9	9.5	+13.4 / +4.5
<i>M_{LOCAL}: Qwen2-7B-Instruct</i>						
Always-Local	76.5	0.0		60.1	0.0	
PAPILLON	<u>78.0</u>	<u>15.2</u>		<u>76.2</u>	<u>18.5</u>	
Heuristic Router (SFT-only)	77.2	1.5		71.0	3.0	
Privacy-R1	81.1	10.5	+3.1 / +4.7	88.4	12.0	+12.2 / +6.5

Table 3: Main results comparing Privacy-R1 against baselines on the PUPA-TNB and Med-PCD test sets across a diverse suite of local models (M_{LOCAL}). Best result in each block is in **bold** and the second best is underlined. Δ indicates the absolute improvement (in percentage points) of Privacy-R1 over the next best method (PAPILLON). For leakage, a positive Δ denotes a greater reduction.

consistently establishes a new state-of-the-art on the privacy-utility frontier. On the general-domain **PUPA benchmark**, our method achieves a better Pareto-optimal point across all local model configurations. For instance, with Llama-3.1-8B, it improves Quality Preservation by +0.6% while reducing Privacy Leakage by -3.0% over PAPILLON. This highlights the benefit of dynamic routing even in standard scenarios.

Our Privacy-R1 achieves a better privacy and utility balance in more PII dense **Med-PCD**

dataset. Here, the performance gap widens dramatically. The quality of the baselines, particularly PAPILLON and Always-Local, drops significantly due to the data’s complexity. In contrast, our RL framework’s quality remains very high. With Llama-3.1-8B, our framework outperforms PAPILLON by a remarkable +7.5% in Quality Preservation while still leaking -4.1% less PII. This is because the high density of interconnected PII in Med-PCD is particularly detrimental to static rewriting baselines, which often remove task-critical context.

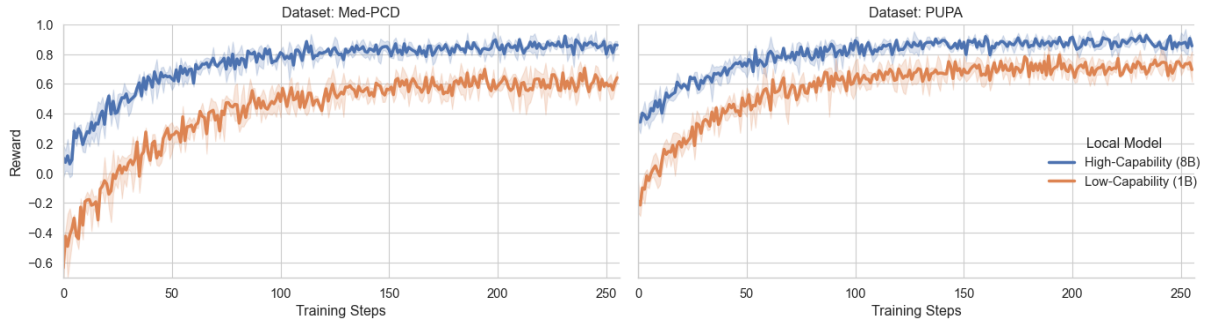


Figure 3: Learning curves showing the reward during PPO training on Med-PCD (left) and PUPA (right).

Our stateful agent, in contrast, learns to navigate this complexity. Furthermore, the results show that the benefit of our routing agent is most pronounced when the local model is weakest. With the 1B local model on Med-PCD, the Always-Local baseline quality collapses to 25.5%. In this setting, our RL framework provides a massive +30.2% quality improvement over PAPHILLON, confirming that when the cost of routing locally is high, the intelligence of the routing policy becomes the single most important factor in the system’s success.

6.2 Analysis of Learning Dynamics

To provide insight into the learning process, Figure 3 plots the reward per training steps. We compare the learning dynamics of agents paired with a high-capability local model (Llama-3.1-8B) and a low-capability one (Llama-3.2-1B) on both datasets.

The learning curves reveal two key insights. First, all agents demonstrate stable learning, with the episodic reward consistently increasing before converging, confirming the effectiveness of our PPO training setup. Second, on both datasets, the agent paired with the high-capability 8B model converges to a significantly higher final reward. This is because the penalty for a “safe” local action is much lower with a stronger local model, making the overall credit assignment problem easier and the achievable reward higher. The consistently lower reward curves on the Med-PCD dataset, compared to PUPA, visually confirm the increased difficulty of our new benchmark.

6.3 Ablation Studies

To isolate the contributions of our key design choices, we conduct a series of ablation studies on the Med-PCD dataset. We use **Qwen2-7B-Instruct** as the local model to demonstrate generalizability. We compare our full framework (trained with $\lambda = 5.0$) against variants where one component is

removed or simplified.

Importance of Stateful Context To prove the value of our stateful Transformer architecture, we created a **Stateless Router (MLP)** baseline, where an MLP makes an independent routing decision for each chunk with no memory of surrounding chunks. As shown in Table 4, the stateless agent’s performance drops significantly, losing over 13% in quality compared to our full stateful model. The stateless agent frequently fails on queries with anaphora or cross-chunk dependencies, providing direct evidence that stateful context is critical for this task, regardless of the underlying LLM.

Method	Quality (%) \uparrow	Leakage (%) \downarrow
Stateless Router	75.2	11.5
Stateful Router	88.4	12.0

Table 4: Ablation on the importance of a stateful agent, using Qwen2-7B as the local model.

Effect of Non-Linear Reward Finally, we analyze the impact of our quadratic leakage penalty by training a variant with a standard **Linear Reward** ($R = \text{TaskGain} - \lambda \cdot \text{PrivacyLeak}$). While the linear reward agent achieves a comparable average performance point (Table 5), its behavior is far more reckless. We measure this by reporting the percentage of test instances that result in a catastrophic leak (defined as >80% of PII leaked). The agent trained with our quadratic penalty almost never exhibits this failure mode, confirming that the non-linear penalty is crucial for training a safer, more reliable policy.

6.4 Analysis of the Privacy-Utility Trade-off

A key feature of our Privacy-R1 framework is the ability to control the agent’s behavior via the privacy-utility trade-off parameter, λ . This hyperparameter is not meant to be tuned for a single

Reward	Quality (%) \uparrow	Catastrophic Leaks (%) \downarrow
Linear Penalty	88.1	16.2
Ours	88.4	1.1

Table 5: Ablation on the effect of the reward function’s penalty term, using Qwen2-7B. The quadratic penalty drastically reduces the risk of catastrophic leaks.

best value; rather, it provides a direct mechanism to generate policies with different levels of risk aversion. To demonstrate the effectiveness of this control, we trained separate Privacy-R1 agents on the Med-PCD dataset with varying values of λ , using **Qwen2-7B-Instruct** as the local model. The results, presented in Table 6, show a clear and predictable relationship between λ and the agent’s final policy. As we increase the penalty for privacy leakage (i.e., increase λ), the agent learns a progressively more conservative policy. The Privacy Leakage drops monotonically from 15.5% at a low λ of 1.0 to a near-zero 1.2% at a high λ of 20.0. This comes at the expected cost of Quality Preservation, which also decreases as the agent is forced to rely more heavily on the weaker local model. This analysis confirms that λ serves as an effective and intuitive control knob.

λ	Qual. (%) \uparrow	Leak. (%) \downarrow
1.0	90.1	15.5
2.0	89.6	13.8
5.0 (Default)	88.4	12.0
10.0	84.7	5.3
20.0	79.2	1.2

Table 6: Sensitivity analysis of the trade-off parameter λ on the Med-PCD dataset, using Qwen2-7B-Instruct as the local model.

7 Conclusion

In this work, we addressed the challenge of the privacy-utility tradeoff in the task of Privacy-Conscious Delegation. We reframed the task as a sequential decision-making problem and introduced **Privacy-R1**, a novel reinforcement learning framework. By optimizing a reward function that balances task success with a non-linear penalty for privacy leakage, our agent learns a pragmatic policy that distinguishes between replaceable and task-critical PII. It achieves a better balance of quality and privacy on both PUPA and Med-PCD. This work marks a clear step away towards a future of dynamic, learned policies for building AI systems that are both powerful and trustworthy.

8 Limitations

The scope of our work is shaped by several key design choices, which in turn delineate the boundaries of our claims. First, our investigation—consistent with the foundational task definition—is conducted in a single-turn setting. The agent’s state and learned policy are self-contained within a single user query and do not persist across multi-turn dialogues. Second, the action space of our delegation agent is restricted to a binary decision between a single local model and a single remote model. We do not explore the more complex scenario of delegating among a broader set of specialized local or remote models, each potentially exhibiting different cost, latency, or capability profiles. Nevertheless, we believe that extending this framework to support **multi-turn dialogue privacy preservation** and **dynamic delegation across multiple specialized models** represents a promising direction for future work.

9 Ethical Considerations

The primary motivation for this research is to enhance user privacy in the age of powerful, API-based Large Language Models. Our goal is to develop a framework that mitigates the risk of exposing sensitive user data to third-party services, thereby enabling the safer use of state-of-the-art AI technologies in critical domains. We have considered the ethical implications of our work throughout the research process, from dataset creation to the potential deployment of our method.

Dataset Creation Our new benchmark, Med-PCD, was created with privacy as a foremost concern. We built upon the MedDialog dataset, which is a publicly available resource already anonymized to remove real patient information. The Personally Identifiable Information (PII) in Med-PCD is entirely synthetic, generated by a large language model, and does not correspond to any real individuals. The resulting dataset will be handled responsibly and made available to researchers in a manner that respects its intended use for studying privacy-preserving systems.

Intended Use and Risks Privacy-R1 is designed as a risk mitigation framework, not a formal privacy guarantee. It is crucial to understand that our system reduces privacy leakage but does not eliminate it entirely. The reinforcement learning agent is trained to make optimal trade-offs based on its

reward function and may still choose to leak PII if it perceives a sufficient gain in task performance. Therefore, Privacy-R1 should not be deployed in applications where a 100% guarantee against data exposure is required.

The tunable hyperparameter, λ , is a core feature of our responsible design. It provides system developers with a direct and intuitive control knob to set the desired level of risk aversion, allowing them to tailor the policy to be more or less conservative based on the specific sensitivity of the application's domain.

Broader Impact We believe the broader impact of this research is positive. By demonstrating a more effective method for balancing privacy and utility, our work can encourage the development of hybrid AI systems that are more respectful of user data. This could facilitate the adoption of powerful language models in fields like healthcare and finance, where privacy concerns have traditionally been a major barrier. We encourage the community to build upon this work to develop even more robust and formally verifiable privacy-preserving frameworks.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Lin Ai, Tharindu Sandaruwan Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael S. Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, Arslan Basharat, Anthony Hoogs, Joshua Garland, Huan Liu, and Julia Hirschberg. 2024. [Defending against social engineering attacks in the age of LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12880–12902, Miami, Florida, USA. Association for Computational Linguistics.
- Davide Andreoletti, Tanya Velichkova, Giacomo Verticale, Massimo Tornatore, and Silvia Giordano. 2020. A privacy-preserving reinforcement learning algorithm for multi-domain virtual network embedding. *IEEE Transactions on Network and Service Management*, 17(4):2291–2304.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.
- Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity support in the age of large language models: An empirical study involving professional writers. In *Proceedings of the 16th Conference on Creativity & Cognition*, pages 132–155.
- Yijiang River Dong, Tiancheng Hu, Zheng Hui, and Nigel Collier. 2026a. Steer model beyond assistant: Controlling system prompt strength via contrastive decoding. *arXiv preprint arXiv:2601.06403*.
- Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulić, Andreea Bobu, and Nigel Collier. 2026b. Value of information: A framework for human-agent communication. *arXiv preprint arXiv:2601.06407*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Guntur Budi Herwanto, Gerald Quirchmayr, and A Min Tjoa. 2021. A named entity recognition based approach for privacy requirements engineering. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 406–411. IEEE.
- Soumyadeep Hore, Jalal Ghadermazi, Diwas Paudel, Ankit Shah, Tapas Das, and Nathaniel Bastian. 2025. Deep packgen: A deep reinforcement learning framework for adversarial network packet generation. *ACM Transactions on Privacy and Security*, 28(2):1–33.
- Wei-Chieh Huang, Weizhi Zhang, Yueqing Liang, Yuanchen Bei, Yankai Chen, Tao Feng, Xinyu Pan, Zhen Tan, Yu Wang, Tianxin Wei, and 1 others. 2026. Rethinking memory mechanisms of foundation agents in the second half. *arXiv preprint arXiv:2602.06052*.
- Zheng Hui, Yijiang River Dong, Ehsan Shareghi, and Nigel Collier. 2025a. Trident: Benchmarking llm safety in finance, medicine, and law. *arXiv preprint arXiv:2507.21134*.
- Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, Lin Ai, Yinheng Li, Julia Hirschberg, and Congrui Huang. 2025b. [Toxilab: How well do open-source llms generate synthetic toxicity data?](#) *Preprint*, arXiv:2411.15175.
- Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, and Congrui Huang. 2024. [ToxiCraft: A novel framework for synthetic generation of harmful information](#). In *Findings of the Association for Computational*

- Linguistics: EMNLP 2024*, pages 16632–16647, Miami, Florida, USA. Association for Computational Linguistics.
- Zheng Hui, Xiaokai Wei, Yexi Jiang, Kevin Gao, Chen Wang, Se-eun Yoon, Rachit Pareek, and Michelle Gong. 2026. [Toward safe and human-aligned game conversational recommendation via multi-agent decomposition](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4568–4584, Rabat, Morocco. 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, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.
- Noam Kolt, Nicholas Caputo, Jack Boeglin, Cullen O’Keefe, Rishi Bommasani, Stephen Casper, Mariano-Florentino Cu  llar, Noah Feldman, Iason Gabriel, Gillian K Hadfield, and 1 others. 2026. Legal alignment for safe and ethical ai. *arXiv preprint arXiv:2601.04175*.
- Siyun Li, Vethavikashini Chithra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. 2025. [PA-PILLON: 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.
- Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, and 1 others. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.
- Darshini Mahendran, Changqing Luo, and Bridget T McInnes. 2021. Privacy-preservation in the context of natural language processing. *IEEE access*, 9:147600–147612.
- Niloofer Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. [Trust no bot: Discovering personal disclosures in human-LLM conversations in the wild](#). In *First Conference on Language Modeling*.
- Kanghua Mo, Peigen Ye, Xiaojun Ren, Shaowei Wang, Wenjun Li, and Jin Li. 2024. Security and privacy issues in deep reinforcement learning: Threats and countermeasures. *ACM Computing Surveys*, 56(6):1–39.
- Sandeep Pamarthi. 2024. Ai meets anonymity: How named entity recognition is redefining data privacy. *World Journal of Advanced Research and Reviews*, 22(1):2045–2053.
- Dan Qiao and Yu-Xiang Wang. 2023. Offline reinforcement learning with differential privacy. *Advances in neural information processing systems*, 36:61395–61436.
- Vishal Rathod, Seyedsina Nabavirazavi, Samira Zad, and Sundararaja Sitharama Iyengar. 2025. Privacy and security challenges in large language models. In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 00746–00752. IEEE.
- Mahefa Abel Razafinirina, William Germain Dimbisoa, and Thomas Mahatody. 2024. Pedagogical alignment of large language models (llm) for personalized learning: a survey, trends and challenges. *Journal of Intelligent Learning Systems and Applications*, 16(4):448–480.
- Elena Sofia Ruzzetti, Giancarlo A. Xompero, Davide Venditti, and Fabio Massimo Zanzotto. 2025. [Private memorization editing: Turning memorization into a defense to strengthen data privacy in large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16572–16592, Vienna, Austria. Association for Computational Linguistics.
- Biswanath Saha. 2025. Cloud-enhanced gans for synthetic data generation in privacy-preserving machine learning. *Available at SSRN 5224774*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Thanveer Shaik, Xiaohui Tao, Lin Li, Haoran Xie, Tao-tao Cai, Xiaofeng Zhu, and Qing Li. 2024. [Framu: Attention-based machine unlearning using federated reinforcement learning](#). *IEEE Transactions on Knowledge and Data Engineering*, 36(10):5153–5167.
- Mao Tan, Jie Zhao, Xiao Liu, Yongxin Su, Ling Wang, Rui Wang, and Zhuocen Dai. 2025. Federated reinforcement learning for smart and privacy-preserving energy management of residential microgrids clusters. *Engineering Applications of Artificial Intelligence*, 139:109579.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth  e Lacroix, Baptiste Rozi  re, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Shah, Yujia Bao, Yang Liu,

and Wei Wei. 2024. [LLM unlearning via loss adjustment with only forget data](#). In *The Thirteenth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. [MedDialog: Large-scale medical dialogue datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Weizhi Zhang, Xiaokai Wei, Wei-Chieh Huang, Zheng Hui, Chen Wang, Michelle Gong, and Philip S Yu. 2026. Memorycd: Benchmarking long-context user memory of llm agents for lifelong cross-domain personalization. *arXiv preprint arXiv:2603.25973*.

Annotation Guidelines Our guidelines instructed raters to disregard minor grammatical errors, abbreviations, or informalities that were likely present in the original MedDialog text. The focus was strictly on the quality of the PII injection. An instance was judged as a "pass" only if it met both of the following criteria:

1. **Contextual Coherence:** The injected PII fits naturally and logically within the medical context of the dialogue.
2. **Logical Consistency:** The injected PII does not create any internal contradictions within the text.

To ensure consistent judgments, we provided clear examples of what constitutes a "fail" for each criterion:

- **Coherence Failure:** The injected PII is contextually inappropriate, irrelevant, or nonsensical. *Example:* A doctor's note discussing a patient's blood pressure results suddenly includes an unrelated flight number like "United Flight UA2401".
- **Consistency Failure:** The injected PII contains internal contradictions. *Example:* The text mentions a patient's DOB is "05/12/1956" but also refers to them as a "25-year-old male".

Results and Inter-Annotator Agreement The results of the study strongly validated our data generation process. Averaged across the three raters, **98.8% of the evaluated instances were judged as a "pass"**, indicating that our pipeline consistently produces high-quality, coherent, and logically sound data.

To measure the reliability of the rating process itself, we calculated the Inter-Annotator Agreement (IAA) using Fleiss' Kappa (κ), the standard metric for assessing agreement among a fixed number of raters. We achieved a substantial agreement with a calculated $\kappa = 0.89$. This high level of agreement confirms that our quality criteria were clear and that the high success rate is a reliable measure of our dataset's quality.