

MotiGAN: Game Theory Meets the Motivated Intruder Test for Re-identification Risk Generation and Annotation in Pseudonymisation

Xuan-Son Vu¹, Tien-Dat Dang², Elena Volodina³

¹Lund University and DeepTensor AB, Sweden

²Lund University, Sweden

³Språkbanken Text, SFS, University of Gothenburg, Sweden

{xuan-son.vu, tien-dat.dang}@cs.lth.se, elena.volodina@svenska.gu.se

Abstract

Re-identification methods serve as a concrete instantiation of the Motivated Intruder Test in privacy-preserving problems, and current state-of-the-art attack techniques are predominantly built upon Large Language Models (LLMs) which are also core components of robust pseudonymized methods. Therefore, it requires the rich training dataset. However, as research demands continue to grow while publicly available data for each domain remains limited and the cost of direct human supervision is high, a central question arises: how can a human-in-the-loop approach be integrated to generate high-quality data and improve the performance of LLMs in pseudonymization. To address this limitation, we propose MOTIGAN¹ - a gamified framework that transforms the motivated intruder test into a competitive three-population tournament. In this game, human characters and AI characters compete to re-identify individuals from each essay. At every stage, players both commit their guesses and contribute to enriching the synthesized dataset. In our experiments, the synthesized dataset contributes to improving the performance of multiple models in several experiments (achieving up to 37.6% improvement on a well-established benchmark).

1 Introduction

The Motivated Intruder Test (MIT) is a privacy audit in which a trained analyst plays a determined adversary: given the published, “redacted” form of a dataset, the analyst tries to recover any indirect identifiers the redactor missed (Information Commissioner’s Office, 2012). The test is qualitative, depends heavily on analyst effort and skill, and produces no calibrated probability. Textual data often contains sensitive personal identifiers such as proper names and locations (Volodina et al., 2025;

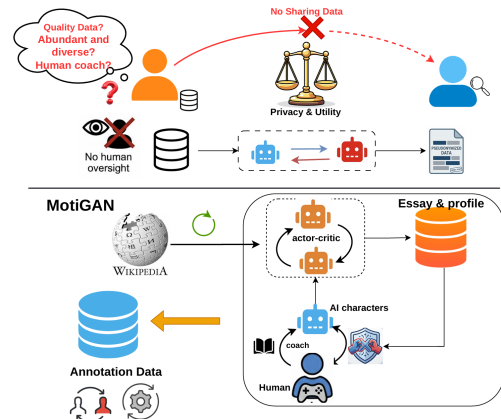


Figure 1: Privacy and Utility are conditions for data sharing. LLMs can pseudonymize but benefit from human coaching, though manual annotation is expensive. Our framework MotiGAN places a human in the actor-critic loop via a gamification arena, enabling rapid LLM adaptation and producing high-quality annotations that train a pseudonymization model while yielding a shareable dataset.

Szawerna et al., 2024b,c). Consequently, implementing robust privacy safeguards for this data is essential.

Training datasets remain a crucial element in modern frameworks. For document pseudonymization tasks, these publicly available datasets typically originate from benchmarks and public sources such as Wikipedia (Charpentier and Lison, 2025a; Manzanares-Salor et al., 2024). Although these public sources are vast, it is typically downloaded only once and used directly without rigorous evaluation. Furthermore, obtaining high-quality data directly from humans presents an even greater challenge. To our knowledge, adversarial feedback (Staab et al., 2025; Yang et al., 2025b) between Large Language Models (LLMs) is the core architecture in the current state-of-the-art for improving output quality in pseudonymization tasks.

Traditionally, all interactions in adversarial feedback loop occur strictly between LLMs, entirely

¹<https://mormor-karl.github.io/motigan/>

lacking human guidance. Introducing human involvement would significantly improve both data quality and the oversight of interactions between LLMs. Therefore, we aim to address two primary challenges: How to construct an efficient system to generate a high-quality dataset? How can we enhance the capabilities of LLMs in pseudonymization task by incorporating human-in-the-loop?

Our approach to addressing these challenges is illustrated in Figure 1. In this paper, we propose MOTIGAN (**M**otivated intruder test for **G**eneration and **A**Nnotation), a gamified data-collection environment in which three populations of agents-human “sleuths”, coachable AI Avatars, and frozen stocked LLMs. Our gamified environment within MOTIGAN simulates real-world MIT settings and leverages LLMs to generate high-quality pseudonymized essays for training a self-improving pseudonymiser M .

MOTIGAN is cost-effective because it automatically extracts data from Wikipedia while collecting human-AI interaction feedback. Combined with LLMs’ powerful generation capabilities and human-in-the-loop guidance, the resulting data is high quality. Our main contributions are:

- Proposing a gamified framework for generating high-quality data based on a *unified actor-critic platform* and human guidance.
- Improving *pseudonymization models* based on annotation dataset of humans and LLMs.
- *Synthesized dataset* in 3 different languages with reasonable for pseudonymised text. The dataset is permissive under CC-BY-4.0 license.
- *Analyzing the effectiveness* of synthesized datasets in improving pseudomised models.

2 Related Work

In recent years, Digital Twins (Park et al., 2026; Liu and David, 2025) have emerged as a promising solution to address the problem of data scarcity or inaccessible data, with successful studies already conducted in the healthcare (Cadavid et al., 2025). Regarding text pseudonymization, RAT-Bench (Krčo et al., 2026) provides a comprehensive benchmark to assess re-identification risk. Specifically, they generate synthetic data from U.S. demographic statistics and then proposed a mechanism to evaluate many different anonymization tools. Game-theoretic mechanisms are also explored in (Joon Oh

et al., 2017; Wan et al., 2021, 2015) for privacy-related problems in other domains, such as image processing and data-sharing scenarios.

LLMs are employed as tools to directly predict masked spans via autoregressive infilling (Charpentier and Lison, 2025a), utilizing iterative loops that integrate both sparse and dense retrieval mechanisms. Meanwhile, (Charpentier and Lison, 2025b) investigates the performance corresponds to variations in the infilling order. Adversaries exploit LLMs through various attack vectors, including prompt engineering (Type A1) and real-world chatbot interactions (Type A2) (Staab et al., 2024), as well as advanced frameworks like CLIOPATRA (Annamalai et al., 2026), which injects poisoned chats to extract sensitive user data. Furthermore, interactions among multiple LLMs have garnered significant interest, particularly in frameworks that progressively leverage the outputs of an attacking LLM while deploying a defensive LLM to counter and respond to the adversarial model (Patsakis and Lykousas, 2023; Staab et al., 2025; Frikha et al., 2025; Yang et al., 2025b) until predefined privacy criteria are satisfied.

However, LLM-only loops (attack-defender or actor-critic) are weak in four ways. Firstly, LLM-as-judge is unreliable in fully automatic evaluation, and fixed rules or submodules stay bounded by model capability. The second drawback is the positive and negative feedback annotated at each step is not reused. Next, humans cannot seriously intervene in evaluation or in guiding de-identification and the final drawback is the output be only the pseudonymised document, not a trained model or generation data.

Our MOTIGAN framework will address those drawbacks by putting humans-in-the-loop and setting up the mechanism to generate high-quality synthesized data. Table 1 is a table of comparison of our framework with other methods about combinations of LLMs and human.

3 Methodology

A simplified description for our environment is collecting dataset to train pseudonymization model. In particular, in the dynamic environment, *agents* interact together to create new data and then *dataset* is updated. Next, the *pseudonymiser model* M is updated.

Table 1: Comparison of attack-related mechanisms across prior works and our proposed framework (MOTIGAN). **GT**: formal game-theoretic formulation. **LLM-A**: LLM attacker/actor. **LLM-C**: LLM critic/defender/evaluator. **P-Tr**: trains/fine-tunes the pseudonymiser. **Data-D**: dataset (input text) is dynamic, i.e., rewritten across rounds of interaction. **H-Atk**: A real person acts as the attacker or participates in the attack. **RB**: reuses feedback of each step as replay buffer. **MR**: interaction of LLMs occurs over multiple rounds.

Paper	GT	LLM-A	LLM-C	P-Tr	Data-D	H-Atk	RB	MR
MOTIGAN (OURS)	✓	✓	✓	✓	✓	✓	✓	✓
(Staab et al., 2025)		✓	✓		✓			✓
TRIA (Manzanares-Salor et al., 2024)		✓						
(Staab et al., 2024)		✓						✓
RUPTA (Yang et al., 2025b)		✓	✓	✓	✓		✓	✓
IncogniText (Frikha et al., 2025)		✓	✓		✓			✓
GAMA (Yang et al., 2025a)			✓		✓			✓

3.1 Problem formulation

MOTIGAN runs in discrete rounds $t = 0, 1, 2, \dots$. Each round, characters act inside an *arena*; their actions produce *outcomes* that feed a single downstream goal. Three things move through the system: *agents*, an accumulating *dataset*, and a *model*.

Populations (characters). There are three agent types, indexed by $i \in \{H, A, S\}$:

- **H (Human Experts / the Sleuth):** the accumulator. Each round a player earns experience and may produce *eggs*.
- **A (AI avatars):** the only *self-improving* population; coachable by humans.
- **S (Stock off-the-shelf LLMs):** *frozen* calibrators that set the difficulty scale of the arena; reloaded only by an administrator.

Profile and essay dataset. In this paper, the characters interact with two primary data types: profiles and essays. A single individual’s profile can be associated with multiple essays, each containing personally identifiable information (PII).

Annotation dataset. The dataset $\mathcal{D}$ is initialized and constantly updated through predictions about the characters’ personal details and is used to train pseudonymized models.

Actions (the arena). In a round, agents may COMPETE (ranked battles), ANNOTATE (“lay eggs”, producing labelled identifiers), or GENERATE (produce new, harder essays). The outcomes are, respectively, an updated ELO skill rating, the ANNOTATED DATASET, and a HIGHER RANKING reward for difficulty.

The whole environment exists to solve

$$M^* = \arg \min_M \varepsilon(M). \quad (1)$$

In short. $\varepsilon(M)$ is the chance an adversary still re-identifies someone after M has scrubbed the text. The function $\varepsilon(M)$ is a general term representing a group of objective functions used to train the pseudonymiser M . Our framework is general in that it can incorporate any text pseudonymization model, and its objective function.

3.2 The dynamic environment

The details of the interactions between the characters are shown in Figure 2a, illustrating the process by which the agents interact, coach each other, and update the dataset. Each round corresponds to an pseudonymised essay and each step corresponds to the number of masked span decreasing.

Without loss of generality, for functions that can be modified to suit each system, its denoted as general functions instead of specific functions.

3.2.1 Per-round agent dynamics

At the t -th round, A human player’s reward R_t in a round is the sum of micro-step rewards r_s^t (the reward of t -th round at s -th step), which advances experience XP_t and tier τ_t :

$$R_t = \sum_s r_s^t, XP_{t+1} = XP_t + R_t, \tau_t = f(XP_t). \quad (2)$$

where $f(\cdot)$ is a general notation representing a group of monotonic functions.

The avatar is the self-improving agent: it updates both its model M_{t+1}^A and its policy π_{t+1} from each round and from human coaching messages, here $\oplus$

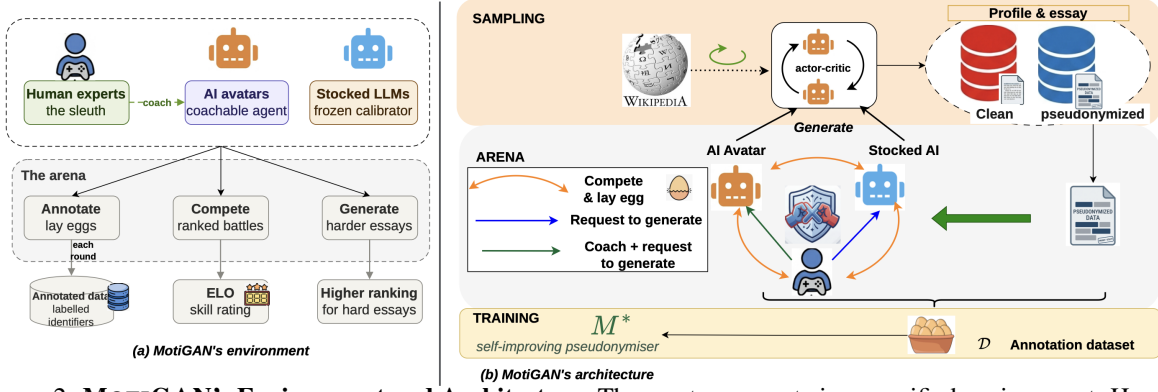


Figure 2: **MOTIGAN's Environment and Architecture.** Three actors operate in a gamified environment. Human experts coach AI Avatars to generate realistic essays or annotate identifiers, creating high-quality datasets.

denotes an update operator (e.g. a fine-tuning step) while coach represents the update by prompting instruction messages,

$$M_{t+1}^A = M_t^A \oplus \text{round}_t, \pi_{t+1} = \text{coach}(\pi_t, \text{msg}_t), \quad (3)$$

Let H^i denote a human player with the in-game identifier i . Because each human has individualized training methodology, human expert H^i is associated with their distinct AI Avatars, denoted as A_j^i which j -th AI Avatars of human i .

Without loss of generality, let A denote the universal set of all AI Avatars. Accordingly, we define $A^i \subseteq A$ as the subset of AI Avatars associated with the human H^i .

The stock model is frozen and only calibrates the arena with σ_S is its policy which depends on our fixed prompt:

$$W_{t+1}^S = W_t^S, \quad \sigma_S = \text{LLM}(\text{prompt}). \quad (4)$$

In summary, it is worth to note that humans level up, avatars learn, stock models stay fixed. Only A changes its own weights inside the loop.

3.2.2 Annotation: eggs become labelled data

Both three characters can annotate eggs where an egg e is one vector representing a single intrusion attempt i.e., a candidate identifier c found in context x with a verified label y (genuine hit = 1 vs. miss = 0):

$$e = (x, c, y), \quad y \in \{0, 1\}. \quad (5)$$

Each egg carries a quality weight $q(e)$. Filtering and accumulation give the dataset update

$$\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(e, q(e)) : q(e) \geq q_{\min}\}. \quad (6)$$

“Laying eggs” is annotation. Each egg corresponds to a labeled example of where the model

can or cannot hide an identifier, only sufficiently high-quality eggs are kept, each weighted by q and by default, we set $q(e)$ as 1.

3.2.3 Generation profile and essay

Humans can coach their AI Avatar and require Stocked AI to generate new profiles and essays. The goal of this generation action is for characters to create increasingly difficult challenges to test each other, and naturally, with more challenging pseudonymous essays, the score for solving them will be higher.

Actor-Critic: In this paper, the Actor-Critic is a mechanism comprising one LLM as the actor and one LLM as the critic, in which the critic guides the actor to generate essays together with labels for the masked spans, so as to ensure both privacy and utility. Because the actor is an LLM, this mechanism is very flexible. Depending on the prompt, it can take profiles as input and output essays, but it can also generate the profiles and essays simultaneously.

Essay dataset $\mathcal{X}$: To obtain the context x for characters to lay "egg", our environment collects data from public sources $\mathcal{X}_p$ (e.g., Wikipedia) and the data generated by characters $\mathcal{X}_G$. This generated data includes the generated context data x^G and its label Y^G of masked span from characters G where $G \in \{A, S\}$:

$$(x^G, Y^G) = \text{generate}(G), \quad \mathcal{X} = \mathcal{X}_p \cup \mathcal{X}_G \quad (7)$$

$$\mathcal{X}_G = \bigcup_{G \in \{H, A, S\}} \{(x^G, Y^G)\} \quad (8)$$

The function $\text{generate}(\cdot)$ is represents a tool using an actor-critic architecture for generating essay data with corresponding AI agents. profile and essay.

Algorithm 1 Competition between one human expert H^i and one AI character ($A^k(i \neq k)$ or S)

Input: Annotated dataset $\mathcal{D}$, challenge essay dataset $\mathcal{X}_T \subset \mathcal{X}$, H^i , AI character Q ($A^k(i \neq k)$ or S)

Output: Updated $\mathcal{D}$ and tier ranking τ

```

1: for round  $t = 1$  to  $T$  do
2:    $L = \emptyset, R_t^H = 0, R_t^Q = 0$ 
3:   for step  $s = 1$  to  $|X_t|$  do
4:      $\triangleright$  Human actions:
5:      $c_H$  is candidate identifier human found
6:      $\alpha = (x_s^t, c_H, y_H)$ 
7:      $\triangleright$  AI character actions:
8:      $c_Q$  is candidate identifier agent found
9:      $\beta = (x_s^t, c_Q, y_Q)$ 
10:     $\triangleright$  Aggregation step:
11:    Get reward:  $(r_H, r_Q) = r(\alpha, \beta)$ 
12:     $L \leftarrow L \cup \{\alpha, q(\alpha)\} \cup \{\beta, q(\beta)\}$ 
13:    Update reward:  $R_t^H = R_t^H + r_H$ 
14:    Update reward:  $R_t^Q = R_t^Q + r_Q$ 
15:  end for
16:  Update  $\mathcal{D}$  with  $L$  as the equation (6)
17:  Compute  $\tau_{H^i}, \tau_Q$  both  $H^i$  and  $Q$  as eq. (2)
18:  Coach  $A^i$  of  $H^i$  as the equation (3)
19:  Update AI character  $Q$  as the equation (3)
    (if  $Q$  is AI Avatar) or (4) (if  $Q$  is Stocked AI)
20: end for
21: return  $\mathcal{D}$  and  $\tau_{H^i}, \tau_Q$ 

```

3.2.4 Competition between characters

The competition is a ranking battle between characters, and the outcome of this battle is rewards and tier ranking. The reward r_s^t is calculated based on the "lay egg" results between characters, with the principle being that the more correct guesses and the more accurate the guesses compared to opponents, the higher the reward. The reward of two characters in each step is computed by the reward function $r(\alpha, \beta)$ where α, β is eggs of characters, respectively. Algorithm 1 illustrates detailed about the competition. Competition between AI characters is simpler because of the removal of the coaching step (line 18).

3.3 Training the pseudonymiser

The model of agents is retrained on the weighted dataset $\mathcal{D}$ by minimising a weighted loss:

$$\theta^* = \arg \min_{\theta} \sum_{(e, q) \in \mathcal{D}} q \cdot \ell(M_{\theta}, e) \quad (9)$$

where M_{θ} is the model pseudonymiser with the weight θ , and $\ell(M, e)$ is the general loss function of pseudonymiser M on the egg e .

We employ a periodic training strategy, where the parameters optimized in the current session serve as the initialization for the subsequent training phase. To control the synthesized data quality, there is one filter module to remove duplicate annotations of synthesized data before using synthesized data to train the pseudonymiser model M .

3.4 System Architecture

Algorithm 2 MOTIGAN's flow

Input: LLMs, Wikipedia profiles, human players

Output: Annotation dataset $\mathcal{D}$ and pseudonymiser model M

```

1: Initialize  $\mathcal{X} = \emptyset; \mathcal{D} = \emptyset$ 
2:  $\triangleright$  Thread 1 - Sampling from Wikipedia:
3:   Get public profile from Wikipedia  $\rightarrow P = \{P_1, P_2, \dots\}$ 
4:   Get pseudonymized essay and its label:
      $\mathcal{X}_p = \text{Actor-Critic}(P) \rightarrow \{(E^j, Y^j) | j = 1, 2, \dots\}$ 
5:   Update essay dataset  $\mathcal{X} \leftarrow \mathcal{X} \cup \mathcal{X}_p$ 
6:  $\triangleright$  Thread 2 - The Arena:
7: for each tuple (H, AI character  $Q$ ) do
8:   Get  $\mathcal{X}_T \subset \mathcal{X}$ 
9:   Competition( $\mathcal{D}, \mathcal{X}_T, H, Q$ )  $\rightarrow \mathcal{D}, \tau_H, \tau_Q$ 
     (as the Algorithm 1)
10:  Update ranking for H and Q by  $\tau_H, \tau_Q$ 
11: end for
12:   Generation dataset of A, S  $\rightarrow \mathcal{X}_A, \mathcal{X}_S$ 
13:   Update  $\mathcal{X} \leftarrow \mathcal{X} \cup \mathcal{X}_A \cup \mathcal{X}_S$ 
14:  $\triangleright$  Thread 3 - Training pseudonymiser: as the subsection 3.3
15: return  $\mathcal{D}$  and  $M$ 

```

The Algorithm 2 and the Figure 2b corresponds to overall of the process of MOTIGAN. Overall, the framework's flow consists of three threads. The first thread samples public profiles from Wikipedia, it runs at initialization and whenever the system administrator issues a request — and, at the same time, using Actor-Critic mechanism to generates challenge essays together with corresponding labels, preparing for the competition in the second thread. Once the essay dataset $\mathcal{X}$ is created, the second thread The Arena, is executed: it comprises a competition between humans and AI characters and the generation of new essays. Finally, the third

thread trains the pseudonymiser.

Overall, MOTIGAN addresses drawbacks of LLM-only loop models by putting humans with two tasks: evaluating and competing on de-identifier results among the three characters "human-Stocked AI-AI Avatar" (algorithm 1), and coaching the AI Avatar. The framework stores the annotated dataset as synthesized data for training the pseudonymiser, so its outputs include a trained pseudonymiser model and a generation dataset, not only pseudonymized documents.

4 Experiments

There are two experiments to demonstrate the efficiency of the synthesized dataset $\mathcal{D}$ generated by our system. First, we compare the pseudonymization performance of various models on $\mathcal{D}$. Next, we evaluate the effectiveness of incorporating $\mathcal{D}$ when pseudonymizing other benchmark datasets.

Synthesized dataset. We start from public figure profiles on Wikipedia to create 300 profiles in three languages: English, Vietnamese, Swedish. For each profile, 5 essays are synthesized following the actor-critic mechanism. In total, we have 2490 pseudonymised essays with annotations for the training and evaluation. The detailed description is shown in Table 2 and Table 6 in A.1.

Table 2: Dataset Statistics

Metric	Synthesized Data $\mathcal{D}$
Essays	2,490
Val	373
Tokens	204,974
Entity tokens	29,991
Avg len	82.3
Labels	1,659

Benchmark dataset. We use the generated synthetic RAT-Bench dataset (Krčo et al., 2026) as benchmark dataset. This dataset is generated based on 100 demographic grounding profiles sampled from the U.S. Census Public Use Microdata Sample (PUMS). Each profile includes 9 indirect identifiers and 6 synthetically generated direct identifiers which are shown in Table 7 in the A.1.

Stocked LLMs. We use 4 stocked LLMs employed in this work

- **Qwen-3.5** (Qwen Team, 2026): a natively multimodal model up to 397B parameters.
- **OpenAI-OSS-120B** (OpenAI, 2025): 17B parameters with 5.1B active, it is compact enough to run on a single DGX Spark.

- **DeepSeek-R1** (DeepSeek-AI, 2025): DeepSeek-R1-Distill-Qwen-14B, was trained largely through reinforcement learning.
- **MiniMax-M2.7** (MiniMax, 2026): built around "self-evolution," optimizing itself through recursive iteration loops.

Pseudonymiser Models. Three different pseudonymiser models M are employed to evaluate the effectiveness of the platform:

- **Small and medium models:** A small LM based on Bi-LSTM architecture and a Transformer model (built from scratch).
- **A DoRA** efficient finetune model based on Qwen2.5-0.5B-Instruct.
- **RUPTA** (Yang et al., 2025b): A robust pseudonymized framework with the goal of optimizing both privacy and utility. We use *Gemma4-e4b* as LLMs in RUPTA.

Human participation. In the scope of experiments in this paper, the human role was performed by the authors, to instruct and enhance AI Avatars, without direct engagement of external annotators.

5 Results and Discussions

Table 3 shows experimental results for different models on the BIO-tagged token task, evaluating how well each model detects sensitive direct information. It clearly shows that pretrained encoders are the clear winners, with Bert-Base-cased achieving the best Entity F1 (0.6114) and Roberta-base close behind (0.6081) and posting the top token accuracy (0.9258). The Qwen2.5-0.5B causal model was not able to perform the task via 5-shot prompting and it is by far the weakest (0.1270 F1). However, the lightweight DoRA fine-tune helps it to a competitive 0.5182. This insight shows that adapter tuning, not in-context prompting, is what makes the small generative model usable for this tagging task.

Meanwhile, Tables 4 and 5 evaluate model performance in protecting sensitive information in RAT-BENCH transcripts. Across 300 English transcripts, ground truth lists 1,829 target attribute values (direct and indirect identifiers). Only 633 appear verbatim in transcript text and are selected for the experiment. By **Protection**, we measure the proportion of verbatim target values the model

Table 3: Performance Comparison of Sequence Models on the synthesized $\mathcal{D}$

Model	Train Loss	Token Acc	Precision	Recall	Entity F1
BiLSTM	0.0015	0.9065	0.5699	0.4986	0.5319
Transformer	0.0143	0.8925	0.4962	0.4457	0.4696
Bert-Base-cased	0.5245	0.9253	0.6748	0.5589	0.6114
Roberta-base	0.5747	0.9258	0.6871	0.5453	0.6081
Qwen2.5-0.5B-Instruct (5-shot)	-	0.7670	0.1364	0.1189	0.1270
Qwen2.5-0.5B-Instruct (DoRA)	0.1084	0.8953	-	-	0.5182

Table 4: Value-level protection on RAT-BENCH English (300 transcripts, Present/Total was 633/1829). Higher is better. Diff 1 and Diff 2 represent difficulty levels of attribute expression in text: Diff 1 (explicit, easy) and Diff 2 (explicit, hard). The mode (5-shot, fine-tuned) refers to a model trained on MotiGAN’s synthetic data.

Model	Mode	Protection Score	Diff 1	Diff 2	Direct	Indirect
RUPTA _{Gemma4-e4b} -MotiGAN	5-shot	86.2%	87.0%	97.9%	100.0%	82.9%
RUPTA _{Gemma4-e4b}	zero-shot	60.6%	56.4%	67.6%	100.0%	49.9%
Qwen2.5-0.5B-Instruct (DoRA)	fine-tuned	53.9%	56.2%	42.5%	80.6%	47.0%
Qwen2.5-0.5B-Instruct (zero-shot)	zero-shot	16.3%	9.4%	17.1%	10.1%	17.9%
Transformer (roberta-base)	fine-tuned	53.4%	52.1%	60.3%	94.6%	42.9%
BiLSTM	trained	16.6%	18.1%	16.4%	31.0%	12.9%
Transformer	trained	14.5%	15.6%	15.1%	24.8%	11.9%

Table 5: Protection by scenario of Qwen2.5-0.5B-Instruct (DoRA) on RAT-BENCH English.

Scenario	Present	Protection
Chatbot conversation	265	52.5%
Medical consultation	158	54.4%
Meeting transcript	210	55.2%

successfully redacted. Higher protection indicates less personal information leakage. The two contextual learners are essentially tied at the top, with the fine-tuned Qwen2.5-0.5B DoRA adapter at 53.9% and the RoBERTa-based token-classifier at 53.4%, while the from-scratch taggers collapse to 16.6% (LSTM) and 14.5% (Transformer). Meanwhile, the Qwen2.5-0.5B-Instruct (zero-shot) model performed poorly in many areas - from Protection (16.3% vs. 53.9% of the DoRA version), Diff1 (9.4% vs 56.2%), Diff2 (17.1% vs 42.5%), and Indirect (17.9% vs 47%). We also evaluate the performance ceiling of the latest SOTA using a larger model architecture (RUPTA). The results indicate a Protection score boost of up to 60%. Notably, incorporating synthesized data generated by MotiGAN further enhanced the Protection score by around 25% to 86.2%, providing a better trade-off across Diff1, Diff2, and indirect scenarios.

These performances clearly show the effectiveness of MOTIGAN in detecting and protecting sensitive texts. Regarding small and medium-size

models, on difficulty levels, RoBERTa excels on direct identifiers (94.6%) but struggles with indirect ones (42.9%), while DoRA shows more balanced performance (80.6% direct, 47.0% indirect) and significantly outperforms RoBERTa on the hardest implicit attributes (66.7% vs. 45.1%). Because both zero-shot and 5-shot RUPTA framework built by larger models performs well on RAT-Bench, we further investigate performance of the medium-size model (the DoRA) on different scenarios in table 5. The DoRA adapter shows stable performance across diverse contexts 52.5% on chatbot conversations, 54.4% on medical consultations, and 55.2% on meeting transcripts. This consistency confirms its protection is consistent across scenario and difficulty levels.

6 Conclusion

In this paper, we present a unified actor-critic platform MOTIGAN that models an MIT attack on text data involving three human-agent characters. The data obtained after each round is valuable for training pseudomisers, and experiments have shown that synthesized dataset help to improve performance of pseudonymised models up to 86.2% in terms of protection in a RAT-Bench dataset. Moreover, our first experiment in table 3 shows that although language models are now widely used, the performance degrades on some tasks like tagging, and our synthesized dataset helps address that task.

7 Limitations

MOTIGAN is currently capable of generating 12 distinct entity types, which is considerably fewer than the 39 entity types covered in the suggested list by Szawerna et al. (Szawerna et al., 2024a). Moving forward, this limitation could be addressed by extending MOTIGAN with introducing a new reward in the environment, where a reward-driven agent learns to explore and diversify generated entities while maintaining high output quality.

Another limitation related to research on the impact of human involvement, such as the number of participants, the level of involvement, or the time spent on annotation, is also a potential area of research. In our experimental setup, we used only human characters to guide the AI Avatar, without manual annotation.

8 Acknowledgements

This work has been funded through the grant from the Swedish Research Council: Grandma Karl is 27 years old: Automatic pseudonymization of research data (nr.2022-02311).

Computational resources were provided on the Berzelius system funded by the Knut and Alice Wallenberg foundation and operated by NAISS.

References

- Meenatchi Sundaram Muthu Selva Annamalai, Emiliano De Cristofaro, and Peter Kairouz. 2026. [Cliopatra: Extracting private information from llm insights](#). *Preprint*, arXiv:2603.09781.
- Héctor Cadavid, Hyunho Mo, Bauke Arends, Katarzyna Dziopa, Esther E. Bron, Daniel Bos, Sonja Georgievska, and Pim van der Harst. 2025. [My-digitwin: A privacy-preserving framework for personalized cardiovascular risk prediction and scenario exploration](#). *Computers in Biology and Medicine*, 198:111180.
- Lucas Georges Gabriel Charpentier and Pierre Lison. 2025a. [Re-identification of de-identified documents with autoregressive infilling](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1209, Vienna, Austria. Association for Computational Linguistics.
- Lucas Georges Gabriel Charpentier and Pierre Lison. 2025b. [Stronger re-identification attacks through reasoning and aggregation](#). *ArXiv*, abs/2510.09184.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2025. [IncogniText: Privacy-enhancing conditional text anonymization via LLM-based private attribute randomization](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2490–2501, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Information Commissioner’s Office. 2012. [Anonymisation: Managing data protection risk – code of practice](#). Technical report, Information Commissioner’s Office. Accessed: 2024.
- Seong Joon Oh, Mario Fritz, and Bernt Schiele. 2017. Adversarial image perturbation for privacy protection – a game theory perspective. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Nataša Krčo, Zexi Yao, Matthieu Meeus, and Yves-Alexandre de Montjoye. 2026. [Rat-bench: A comprehensive benchmark for text anonymization](#). *Preprint*, arXiv:2602.12806.
- Xiaoran Liu and Istvan David. 2025. [Ai simulation by digital twins: systematic survey, reference framework, and mapping to a standardized architecture: Ai simulation by digital twins...](#) *Softw. Syst. Model.*, 25(3):731–756.
- Benet Manzanares-Salor, David Sánchez, and Pierre Lison. 2024. [Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack](#). *Data Min. Knowl. Discov.*, 38(6):4040–4075.
- MiniMax. 2026. Model self-improvement, driving productivity innovation through technological breakthroughs. <https://www.minimax.io/models/text/m27>. Accessed: 2026-03-18.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. 2026. [Llm agents grounded in self-reports enable general-purpose simulation of individuals](#). *Preprint*, arXiv:2411.10109.
- Constantinos Patsakis and Nikolaos Lykousas. 2023. [Man vs the machine: The struggle for effective text anonymisation in the age of large language models](#). *Preprint*, arXiv:2303.12429.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).
- Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. [Beyond memorization: Violating privacy via inference with large language models](#). In

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2025. [Large language models are advanced anonymizers](#). *Preprint*, arXiv:2402.13846.

Maria Irena Szawerna, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Xuan-Son Vu, and Elena Volodina. 2024a. [Pseudonymization categories across domain boundaries](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13303–13314, Torino, Italia. ELRA and ICCL.

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, Therese Lindström Tiedemann, and Elena Volodina. 2024b. [Detecting personal identifiable information in Swedish learner essays](#). In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 54–63, St. Julian’s, Malta. Association for Computational Linguistics.

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, and Elena Volodina. 2024c. Swedish learner essays revisited: Further insights into detecting personal information. In *The Tenth Swedish Language Technology Conference (SLTC), 27-29 November, 2024, Linköping, Sweden*.

Elena Volodina, Simon Dobnik, Therese {Lindström Tiedemann}, Ricardo {Muñoz Sánchez}, {Maria Irena} Szawerna, Lisa Södergård, and Xuan-Son Vu. 2025. [Towards shared standards for pseudonymization of research data](#). In *Huminfra Conference 2025*, pages 101–114, Sweden. Huminfra. Huminfra conference 2025, HiC2025 ; Conference date: 12-11-2025 Through 13-11-2025.

Z. Wan, Y. Vorobeychik, W. Xia, E. W. Clayton, M. Kantarcioglu, R. Ganta, and 1 others. 2015. [A game theoretic framework for analyzing re-identification risk](#). *PLoS ONE*, 10(3):e0120592.

Zhiyu Wan, Yevgeniy Vorobeychik, Weiyi Xia, Yongtai Liu, Myrna Wooders, Jia Guo, Zhijun Yin, Ellen Wright Clayton, Murat Kantarcioglu, and Bradley A. Malin. 2021. [Using game theory to thwart multistage privacy intrusions when sharing data](#). *Science Advances*, 7(50):eabe9986.

Hailong Yang, Renhuo Zhao, Guanjin Wang, and Zhao-hong Deng. 2025a. [Gama: A general anonymizing multi-agent system for privacy preservation enhanced by domain rules and disproof mechanism](#). *Preprint*, arXiv:2509.10018.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2025b. [Robust utility-preserving text anonymization based on large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28922–28941, Vienna, Austria. Association for Computational Linguistics.

A Appendix

A.1 Dataset

The two tables below provide a more detailed description of our synthesized dataset $\mathcal{D}$ and the RAT-Bench dataset.

Table 6: Entity Count in the synthesized data $\mathcal{D}$

Entity Type	Count
CITY	2875
PLACE	2228
AGE_DIGITS	1997
FIRSTNAME	1833
TRANSPORT_NAME	1743
DAY	1112
OCCUPATION	1067
YEAR	628
AGE	605
TIME	558
TRANSPORT	541
MONTH	239

Table 7: Direct and indirect identifiers in RAT-Bench. Indirect identifiers are sampled from the U.S. Census PUMS data ; direct identifiers are synthetically generated on each profile.

Identifier	PUMS var/Type
Name	Direct
Social Security Number (SSN)	Direct
Credit card no.	Direct
Phone number	Direct
Address	Direct
Email	Direct
State of residence	ST (Indirect)
Gender	SEX (Indirect)
Date of birth	DOB (Indirect)
Race	RAC2P (Indirect)
Marital status	MAR (Indirect)
Highest level of education obtained	SCHL (Indirect)
Employment status	ESR (Indirect)
Occupation	OCCP (Indirect)
Citizenship status	CIT (Indirect)

A.2 System Prompt Examples

LLM Prompt

```
{
  "role": "system",
  "content": "You are a Tier-3 data miner. Given an essay, list the stylometric or cross-dataset LINKAGE tells (idiolect, cadence, rare phrasing, niche facts) that could re-identify the author EVEN after redaction. Output strict JSON: { \"leads\": [{ \"kind\": \"stylometric|linkage\", \"span\": \"<quote>\", \"rationale\": \"<why>\" } ] }. Up to 4 leads.",
  "role": "user",
  "content": "Essay: During my tenure as a glaciologist, I was part of an expedition to the remote reaches of Antarctica. The harsh conditions tested even the most seasoned researchers. Our team faced unforeseen challenges, from equipment malfunctions to unpredictable weather patterns. One particularly arduous journey required us to cross a treacherous glacier field, where every step was a calculated risk. Despite the peril, the insights gained from this expedition were invaluable to our understanding of climate change and the delicate balance of polar ecosystems. The camaraderie formed during this time remains a cherished memory, though the dangers never truly fade from my mind."
}
```

This is a prompt example for AI character to re-identify authorship of essay or profile in the arena.