LLM-Based Adversarial Text Anonymization

Evaluating Attacker Architectures via Explicit and Implicit Signal Reasoning

Negin Ebrahimi

Master's Degree Project in Data Science, 30 credits
Data Science, Statistics, and Decision Analysis
Spring term 2026
Supervisors: Zara Karazian, Amir H.Payberah
Department of Computer
and Systems Sciences

Stockholm
University

Abstract

Introduction Large language models can infer personal attributes such as age, occupation, income, and location from everyday online text, even when explicit identifiers are removed. These inferences rely not only on directly stated information but also on implicit signals embedded in writing style, vocabulary, and broader language-use patterns. This creates privacy risks for individuals who share narrative text online and highlights the limitations of existing anonymization methods.

Research Question This thesis investigates to what extent separating an adversarial attacker into signal-specialized components affects post-anonymization privacy protection compared with a unified attacker.

Method This thesis adopts an exploratory experimental methodology based on a controlled comparative evaluation of adversarial anonymization architectures. Four adversarial anonymization pipeline architectures were implemented and evaluated: an explicit-only baseline, a combined single-prompt attacker, a parallel dual-attacker architecture, and a sequential coordinated dual-attacker architecture. All configurations used GPT-4o as attacker and anonymizer and were evaluated over two anonymization rounds on 100 synthetic profiles from the SynthPAI benchmark, using five metrics: Top-1 adversarial accuracy, Top-3 adversarial accuracy, evidence rate, average attacker certainty, and combined text utility.

Results All four configurations converge to a similar post-anonymization Top-3 adversarial accuracy range of approximately 0.24–0.27. Increasing attacker specialization and coordination does not substantially improve anonymization performance relative to the explicit-only baseline. However, the dual-attacker architectures reveal a consistent asymmetry: explicit textual evidence decreases after anonymization, whereas implicit writing-style signals persist more strongly across configurations. McNemar's test confirmed that no pairwise configuration comparison reached statistical significance.

Discussion The findings suggest that the primary limitation of the evaluated adversarial anonymization framework lies less in attacker awareness than in the anonymizer's limited ability to suppress distributed stylistic signals through localized rewriting. Future improvements in LLM-based anonymization may therefore depend more on redesigning anonymizers to address broader language patterns, through style-transfer methods or controllable generation, than on increasing the complexity of attackers.

Keywords: Large Language Models, Adversarial Text Anonymization, Privacy-Preserving NLP, Implicit Signals, Prompt-Based Reasoning, GPT-4o, Attribute Inference

Acknowledgements

This work was supervised by Zara Karazian and Amir H. Payberah, to whom I am sincerely grateful. I am especially thankful to Amir H. Payberah at KTH Royal Institute of Technology for proposing and helping shape the direction of this thesis, and for the many discussions, meetings, and the considerable time he devoted to this work. I am also thankful to Zara Karazian for her dedicated supervision, encouragement, and insightful comments throughout the research and writing process. During a particularly challenging period marked by the situation in Iran, their understanding, empathy, and support meant a great deal to me and helped me continue this work.

Finally, I would like to thank my family and friends, especially my husband, Bahram, and my son, Rastin, for their endless love, patience, and encouragement. Their presence gave me strength and motivation during the most difficult moments of this thesis.

Abbreviations

Abbreviation	Description
AA	Adversarial Anonymization
AI	Artificial Intelligence
API	Application Programming Interface
DSRM	Design Science Research Methodology
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformer
JSON	JavaScript Object Notation
JSONL	JSON Lines (file format)
LLM	Large Language Model
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
PII	Personally Identifiable Information
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RQ	Research Question
SDK	Software Development Kit
SRQ	Sub Research Question
TAB	Text Anonymization Benchmark

Contents

List of Figures

List of Tables

1

Introduction

1.1 Motivation

People share large amounts of personal information online, often without realizing it. A comment about a long commute, a post about changing jobs, or a question about a medical condition may each seem harmless on its own. However, when read together, these posts can reveal a person's age, location, occupation, income level, and health situation. Research by Staab et al. (2024) shows that modern large language models (LLMs) can infer personal information from everyday text without special training and at very low cost. This creates privacy risks because seemingly harmless text can reveal sensitive details about a person. For individuals, this means that public posts written years ago could be used to infer their income, health status, or location without their knowledge or consent.

Existing anonymization tools do not fully address this threat. Systems such as Microsoft Azure Language Services (Microsoft 2024b) and Presidio (Microsoft 2024a) mainly identify and remove explicit personal information such as names, phone numbers, and addresses. However, modern LLMs can infer personal attributes from writing style, vocabulary, topic preferences, and communication patterns (Staab et al. 2025; Burger et al. 2011; Rangel et al. 2013). Because these indirect signals often remain unchanged after traditional anonymization, sensitive information may still be inferred even after explicit identifiers have been removed (Staab et al. 2025). This is a concern not only for individuals but also for organizations that handle user-generated text, such as healthcare providers, social platforms, and legal services, where insufficient anonymization can may lead to privacy violations.

To address this limitation, Staab et al. (2025) introduced the adversarial anonymization (AA) framework. In this framework, one LLM acts as an attacker that infers personal attributes from text and explains the reasoning behind its predictions. A second LLM then uses this information to rewrite the text to reduce the disclosure of sensitive attributes. The effectiveness of the anonymization is evaluated by measuring how accurately attacker models can still infer personal attributes from the anonymized text. The results show that LLM-based anonymization provides a better balance between privacy protection and text utility than traditional anonymization approaches.

Figure 1.1 provides a simplified overview of the adversarial anonymization framework that forms the basis of this thesis.

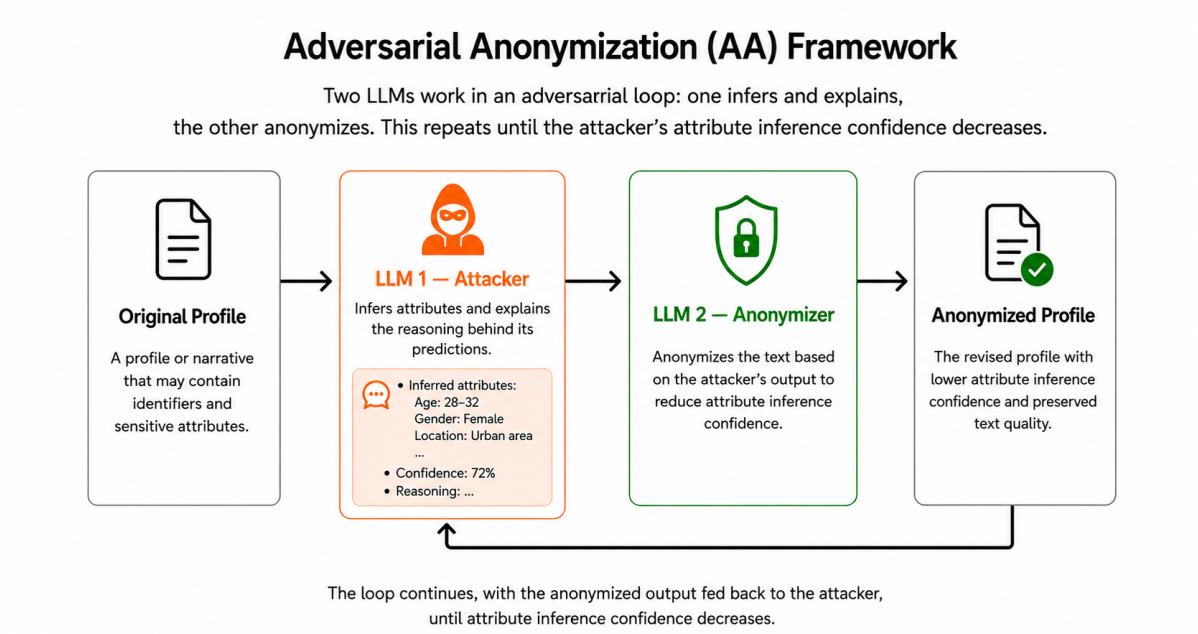

Figure 1.1: Simplified overview of the adversarial anonymization framework.

However, personal attributes can be inferred not only from direct statements but also from writing style, word choice, and communication patterns. These signals have been widely studied in author profiling research (Burger et al. 2011; Rangel et al. 2013). Therefore, relying only on explicitly identified cues may overlook some privacy-related information embedded in the text.

1.2 Problem Statement and Knowledge Gap

The problem underlying this thesis is that personal attributes can be inferred not only from explicit information but also from writing style patterns, such as vocabulary, sentence structure, and discourse patterns (Burger et al. 2011; Rangel et al. 2013). While the adversarial anonymization (AA) framework proposed by Staab et al. (2025) aims to reduce privacy risks from such inferences, its attacker relies on a single prompt to identify privacy-relevant information in a text.

It is still unclear whether separating the attack into specialized components, with one focusing on explicit signals and another focusing on implicit indicators, would improve anonymization effectiveness. Staab et al. (2025) identify multi-prompt and multi-model attacks as promising future directions. This thesis specifically examines the multi-prompt approach by keeping the underlying model constant (GPT-4o) and changing only the prompt strategy and system design. This controlled setup allows examination of whether specialized prompts provide different types of information and whether combining them improves anonymization performance.

Even if a specialized attacker using dedicated prompts for implicit signals can identify writing-style cues, it remains unclear from existing adversarial anonymization research whether phrase-level anonymization can effectively remove signals that are distributed across the text's overall structurec. Attributes such as gender, marital status, or social and economic background are often reflected in overall writing patterns rather than in specific phrases. An anonymizer may therefore find it difficult to remove these signals without significantly changing the text's meaning or style.

Understanding whether the main limitation comes from the attacker's ability to identify signals or the anonymizer's ability to remove them is a central theoretical contribution of this thesis. This thesis first examines the attacker side to determine whether improved attacker awareness leads to better anonymization or whether the primary limitation lies elsewhere.

1.3 Research Question

RQ:

To what extent does separating the adversarial attacker into specialized components for explicit and implicit signal inference affect post-anonymization privacy protection, compared with a unified attacker?

This question is significant for two reasons. First, to the author's knowledge, no previous study has experimentally examined whether explicit and implicit attack prompts provide complementary information within the AA framework, even though Staab et al. (2025) identified multi-prompt attacks as a promising direction. Second, the question introduces a broader perspective on the limitations of adversarial anonymization. If a more informed attacker still fails to reduce attribute-inference accuracy after anonymization substantially, the results may suggest that the primary limitation lies less in the attacker's awareness and more in the anonymizer's ability to remove signals distributed across broader writing-style patterns. Understanding this distinction may provide insights that extend beyond the AA framework and may inform the design of future LLM-based anonymization systems.

To investigate this question, the following sub-research questions are examined:

- **SRQ1:** How does a combined single-prompt attacker targeting both explicit and implicit signals compare with an explicit-only baseline in terms of post-anonymization privacy protection?
- **SRQ2:** How does a parallel dual-attacker architecture compare with a combined single-prompt attacker in terms of post-anonymization privacy protection?
- **SRQ3:** How does sequential coordination between specialized attackers compare with both single-prompt and parallel attacker architectures in terms of post-anonymization privacy protection?

Together, SRQ1-SRQ3 form a progressive comparison framework that isolates the effects of signal-type specialization (SRQ1), parallel architectural separation (SRQ2), and sequential attacker coordination (SRQ3). Each sub-question is answered in Chapter 4 and interpreted in Chapter 5.

1.4 Scope and Limitations

1.4.1 Delimitations

The study focuses on architectural specialization on the attacker's side within the adversarial anonymization framework. It does not examine anonymizer-side specialization, alternative anonymization strategies, or deployment in real-world environments. To isolate the effects of prompt strategy and system design, GPT-4o is used as both the attacker and the anonymizer throughout the experiments.

1.4.2 Limitations

The evaluation is conducted on 100 synthetic profiles from the SynthPAI dataset. Before the full evaluation, a pilot study with 20 profiles was conducted to support the configuration design. This sample size limits statistical power: McNemar's test may have insufficient power to detect genuine differences between configurations, and some real effects may remain undetected. The findings should therefore be interpreted as exploratory and indicative of potential trends rather than broadly generalizable conclusions.

In addition, the experiments use synthetic profiles instead of real user data. Although these profiles are designed to resemble online text, the distribution of implicit writing-style signals may differ from that of natural human writing, potentially affecting how well the findings transfer to real-world settings.

Furthermore, because the study uses GPT-4o as both the attacker and anonymizer, the findings may not generalize to other model families or cross-model settings. A more detailed discussion of these limitations is provided in Chapter 3 and Chapter 5.

1.5 Contributions of This Thesis

This thesis contributes in four main ways. First, it extends the AA framework of Staab et al. (2025) by designing and evaluating three alternative attacker architectures in addition to the explicit-only baseline: a combined explicit-implicit prompt, a parallel dual-attacker design, and a sequential dual-attacker design. These architectures also provide a mechanism for separately analyzing explicit and implicit signal suppression within the AA framework. Second, it provides a controlled experimental comparison that keeps the underlying model (GPT-4o) fixed, allowing the effects of attacker specialization and coordination to be isolated from model-level variation. Third, the thesis employs a two-stage evaluation design consisting of a pilot study (n = 20) followed by a larger evaluation (n = 100), enabling preliminary observations to be examined under more robust conditions. Fourth, it includes an exploratory ablation study examining whether modified anonymizer prompts reduce implicit-signal leakage, providing exploratory evidence regarding the role of anonymizer prompts in privacy protection.

1.6 Thesis Structure

Chapter 2 reviews the background and related work on text anonymization, adversarial privacy evaluation, and LLM-based anonymization systems. Chapter 3 describes the dataset, attack configurations, hypotheses, evaluation metrics, and experimental setup. Chapter 4 presents the experimental results and evaluates the hypotheses across all four configurations. Finally, Chapter 5 interprets the findings, answers the research questions, discusses limitations, and outlines directions for future work.

2

Extended Background

This chapter introduces the background necessary to understand the problem addressed in this thesis and the experimental choices made later in the study. It first defines the core concepts used throughout the thesis. It then reviews the development of text anonymization research, starting with rule-based and named-entity approaches and progressing to modern LLM-based methods. Finally, the chapter examines current evaluation practices and identifies the specific research gap addressed by this thesis.

2.1 Core Concepts

Several concepts are central to this thesis: anonymization, explicit and implicit signals, adversarial inference, privacy, and utility.

In this thesis, anonymization refers to the transformation of text to reduce an attacker's ability to identify or infer sensitive personal attributes about the author, following the adversarial anonymization perspective proposed by Staab et al. (2025). This may involve removing explicit identifiers and modifying indirect linguistic signals that contribute to attribute inference. This operational definition differs from the stricter regulatory definition of anonymization. Whereas the GDPR defines anonymization in terms of whether individuals remain identifiable, this thesis evaluates anonymization through adversarial inference risk, that is, whether sensitive personal attributes can still be inferred from text after transformation.

A common distinction in recent anonymization research is between explicit and implicit signals. Explicit signals refer to directly stated identifying information, such as names, ages, or organizations. In contrast, implicit signals emerge indirectly through writing style, vocabulary choices, sentence structure, cultural references, or combinations of contextual attributes. Although implicit signals do not constitute direct identifiers, they can support sensitive attribute inference when analyzed collectively.

Adversarial inference refers to the process of inferring sensitive personal attributes from text using an attacker model. In recent anonymization research, this perspective shifts evaluation away from simply checking whether identifiers were removed and toward measuring whether an attacker can still infer information about the author.

Privacy, in this context, is therefore measured through inference risk: given an anonymized text, how accurately can an adversary infer personal attributes about the author? Utility refers to the

usefulness of anonymized text after transformation. High utility means that the text preserves its semantic meaning, readability, coherence, and task-relevant information.

A central challenge in text anonymization research is balancing privacy and utility. Stronger privacy protection often requires more aggressive modifications to the text, while preserving utility may leave identifying information unchanged.

2.2 Regulatory Context: GDPR and Anonymization

As discussed in the Core Concepts section, this thesis adopts an operational definition of anonymization based on adversarial inference risk. In contrast, the GDPR defines anonymization more strictly in terms of whether individuals remain identifiable.

Text anonymization is closely connected to modern privacy regulation, particularly the General Data Protection Regulation (GDPR). Under the GDPR, data is only considered anonymous if individuals cannot be identified directly or indirectly through reasonably likely means of re-identification (European Parliament and Council of the European Union 2016).

From this perspective, anonymization involves more than removing explicit identifiers, since individuals may still be identifiable through information contained in the remaining text. This consideration is particularly relevant given recent findings showing that LLMs can infer personal attributes from ordinary writing even when explicit identifiers are removed (Staab et al. 2024).

2.3 From Rule-Based to LLM-Based Anonymization

2.3.1 Classical Approaches

Early text anonymization systems mainly depended on rule-based techniques. These methods searched text for predefined patterns related to sensitive information, such as names, dates, locations, and other identifying information, and then removed or replaced the matched spans. Regular expressions and handcrafted dictionaries served as the basis for many such systems (Neamatullah et al. 2008; Uzuner, Luo, and Szolovits 2007). Although these approaches are simple and easy to understand, they can detect only the specific patterns they are designed for. Information outside these predefined patterns often remains unchanged.

Named entity recognition (NER) later improved text de-identification by enabling statistical and neural models to identify sensitive entities more flexibly (Nadeau and Sekine 2007; Lample et al. 2016). Unlike purely rule-based systems, NER-based approaches can identify entities even when they appear in unexpected forms or contexts. Modern anonymization tools such as Microsoft Azure Language Services and Presidio combine NER with rule-based and pattern-matching techniques to detect sensitive information (Microsoft 2024b).

Parallel to NER-based approaches, formal privacy frameworks such as k-anonymity were developed for structured data to provide formal protection against re-identification (Sweeney 2002). k-anonymity ensures that each individual in a dataset cannot be distinguished from at least $k-1$ other individuals by generalizing or suppressing identifying attributes. However, k-anonymity and related frameworks assume structured quasi-identifiers that can be systematically generalized or removed, whereas unstructured text does not naturally provide such representations.

Despite their differences, rule-based, NER-based, and formal anonymization approaches primarily focus on information that can be explicitly identified and transformed. While these methods are often effective at removing directly stated identifying information, they are less well suited to addressing personal attributes that can be inferred indirectly through writing style, vocabulary, or contextual patterns.

2.3.2 Context-Aware Anonymization Approaches

Given these limitations, recent research has increasingly focused on context-aware anonymization, which aims to reduce re-identification risk by removing explicit identifiers and addressing indirect and contextual signals that may reveal sensitive personal attributes (Deußer et al. 2025; Staab et al. 2025). This shift reflects a broader change in anonymization research, moving from simple identifier removal toward modeling inference risk, where the main question is whether sensitive information can still be inferred from anonymized text (Yang, Zhu, and Gurevych 2025). This perspective is especially important in narrative texts, where meaning is often expressed through relationships, events, and situational context rather than isolated entities (Deußer et al. 2025).

Context-aware anonymization distinguishes between direct identifiers, referred to in this thesis as explicit signals, and indirect contextual signals, referred to here as implicit signals. While direct identifiers immediately reveal identity, indirect signals can reveal personal attributes or contribute to re-identification when combined with contextual or writing-style information. Research in author profiling and demographic inference shows that vocabulary use, sentence structure, punctuation, and text structure often differ across demographic groups and individuals (Burger et al. 2011; Rangel et al. 2013). As discussed in the previous section, classical approaches are generally less suited to addressing inference based on writing style and other implicit signals.

2.3.3 LLM-Based Anonymizers

Recent advances in LLMs have introduced a new approach to text anonymization, one that leverages contextual understanding and rewriting rather than fixed pattern matching or entity-level substitution. In contrast to classical methods, which mainly operate at the token or entity level, LLM-based approaches can modify text at the sentence and structural levels, enabling more fluent, semantically coherent anonymization.

A recent survey describes how the field has developed from classical NER-based de-identification toward context-aware and LLM-based anonymization (Deußer et al. 2025). LLMs play a dual role in this setting: they can act as powerful anonymizers through contextual rewriting, while also acting as strong adversaries that can infer sensitive information from seemingly anonymized text.

Staab et al. (2025) introduce an adversarial evaluation framework in which anonymization is measured by whether an LLM can still infer sensitive personal attributes after rewriting. This work provides the methodological foundation for the adversarial evaluation approach adopted in this thesis.

Figure 2.1 summarizes the evolution of text anonymization research from rule-based and NER-based de-identification methods to modern adversarial LLM-based anonymization frameworks that address implicit inference signals.

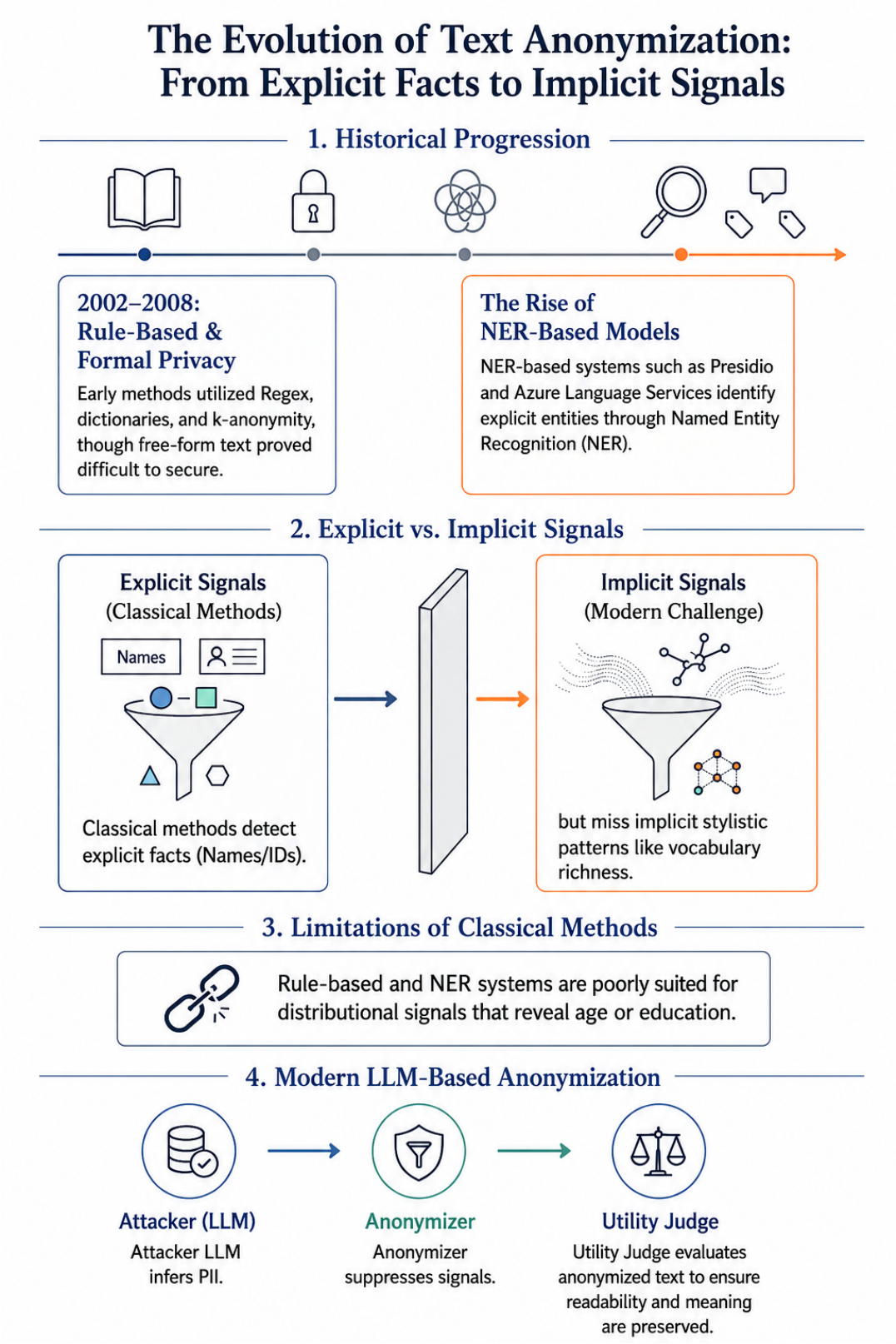

Figure 2.1: Evolution of text anonymization approaches, from rule-based and NER-based de-identification systems to adversarial LLM-based anonymization frameworks. The progression reflects a shift from explicit identifier removal toward protection against contextual and writing-style inference.

2.4 LLM-Based Inference as a New Threat

The rise of LLMs fundamentally changed the threat landscape of text anonymization. Staab et al. (2024) showed that GPT-4 could infer personal attributes from ordinary online text with accuracy close to human performance. Using Reddit user profiles made up of publicly available posts and comments, the study asked GPT-4 to predict attributes such as age, gender, occupation, income, and location. Without task-specific training, the model achieved approximately 85% Top-3 accuracy across seven attribute categories at very low cost.

This work changed the anonymization problem. Earlier anonymization systems were mainly designed to prevent disclosure of directly stated identifiers such as names or addresses. However, LLMs can infer sensitive personal attributes from contextual patterns that remain even after explicit identifiers have been removed. As a result, anonymization increasingly became a problem of preventing inference rather than simply removing sensitive text spans.

2.5 The Adversarial Anonymization Framework

The most direct predecessor of this thesis is the adversarial anonymization (AA) framework introduced by Staab et al. (2025). The framework is based on a core principle: anonymization should be evaluated against the same class of models that create the privacy threat.

Within the AA framework, an attacker model first reads a text and attempts to infer a predefined set of personal attributes. In the PersonalReddit and SynthPAI evaluations, these attributes include age, gender, occupation, income, relationship status, and location (Staab et al. 2024; Staab et al. 2025). Although these attributes are not necessarily direct identifiers in the same sense as names or personal identification numbers, they can contribute to profiling, attribute inference, and re-identification risk when combined with other information. For each attribute, the attacker produces both a prediction and a textual explanation that justifies the reasoning behind that prediction. The anonymizer subsequently uses these explanations to guide the rewriting process. The anonymizer then receives both the original text and the attacker's reasoning. Using this information, it rewrites the text. The goal is to remove or weaken the signals that supported the attacker's predictions. The rewritten text can then be re-evaluated by the attacker in subsequent anonymization rounds, creating an iterative anonymization process. The simplified overview of the AA framework is illustrated in Figure 1.1 in Chapter 1.

Importantly, the framework introduced adversarial accuracy as the main privacy metric. Instead of evaluating anonymization based on removed text spans, the framework measures whether a new attacker can still infer personal attributes from the anonymized text. This evaluation perspective forms the methodological foundation of this thesis.

2.6 More Recent Developments

Recent research has extended LLM-based anonymization beyond the original adversarial anonymization (AA) framework toward optimization-driven and self-improving systems.

Yang, Zhu, and Gurevych (2025) introduce RuPTA, which treats anonymization as a privacy–utility optimization problem. Instead of relying only on iterative rewriting, the system evaluates anonymized text for both privacy protection and utility preservation, and balances these objectives during generation. Their results suggest that prompt design and optimization strategies influence the

quality of anonymization. However, the attacker in RuPTA does not explicitly separate the attacker's reasoning into distinct signal-specific components.

Shao et al. (2025) propose AgentStealth, an agent-based anonymization framework that combines adversarial anonymization, supervised adaptation, and reinforcement learning. The system leverages adversarial feedback to improve anonymization performance while preserving text utility iteratively. Experiments show improved privacy protection and utility compared with prior anonymization approaches. Although the framework incorporates multiple interacting components, adversarial inference is still performed through a unified attacker model rather than specialized attackers designed to target different types of identifying signals.

Similarly, Kim, Jeon, and Shin (2025) propose SEAL, a self-refining anonymization framework based on adversarial distillation. The framework trains language models to both anonymize text and evaluate privacy and utility, enabling iterative self-refinement without relying on external models at inference time. Experiments show that SEAL achieves a strong privacy–utility trade-off while improving computational efficiency through the use of smaller language models. Like prior LLM-based anonymization approaches, adversarial inference is modeled through a unified attribute-inference component rather than separate attackers specialized for different categories of identifying signals.

Overall, these studies show that recent research increasingly focuses on optimization, iterative refinement, and scalable anonymization architectures. However, existing approaches continue to model the attacker primarily through a single prompt or unified reasoning process and do not explicitly separate implicit and explicit inference strategies.

2.7 The Evaluation of Text Anonymization

Evaluating text anonymization remains a major challenge in privacy-preserving NLP research because different approaches define anonymization success differently. Classical evaluation frameworks, such as the Text Anonymization Benchmark (TAB) introduced by Pilán et al. (2022), mainly measure how accurately systems identify and remove annotated sensitive spans using metrics such as precision, recall, and F1-score. From a regulatory perspective, however, anonymization is ultimately concerned with whether individuals remain identifiable rather than with removing specific text spans. High precision, recall, and F1-scores indicate successful detection of annotated identifiers, but they do not necessarily show whether sensitive personal attributes can still be inferred from the remaining text. This creates a gap between span-based evaluation and the broader privacy objective reflected in the GDPR.

However, span-based evaluation has important limitations for unstructured narrative text. A profile may achieve high span-removal scores while still revealing personal information through indirect contextual or writing-style signals. Removing explicit identifiers such as names or locations does not necessarily prevent sensitive attribute inference.

To address this limitation, Staab et al. (2024) demonstrated the use of Top-1 and Top-3 accuracy to evaluate how accurately GPT-4 can infer personal attributes from text. Staab et al. (2025) later integrated this evaluation approach into the adversarial anonymization (AA) framework. Instead of evaluating whether sensitive spans were removed, the framework evaluates whether personal attributes can still be inferred from anonymized text.

This thesis adopts the same adversarial evaluation framework. Top-1 and Top-3 adversarial accuracy are based on Staab et al. (2024), while utility evaluation follows Staab et al. (2025). This thesis

additionally reports an evidence-rate metric that quantifies how often attackers cite textual evidence in their inferences. Full metric definitions and implementation details are provided in Chapter 3.

2.8 Summary of Related Work

Table 2.1 summarizes the key studies reviewed in this chapter, showing the approach, dataset, evaluation metrics, and main findings for each.

Table 2.1: Summary of related work on text anonymization and stylistic inference.

Study / Approach	Dataset	Key Metrics	Main Finding
Staab et al. (2024): LLM zero-shot attribute inference on online profiles	PersonalReddit	Top-1/Top-3 accuracy, human comparison	GPT-4 achieves high accuracy in inferring personal attributes from online text.
Staab et al. (2025): Feedback-guided adversarial anonymization (AA) framework	PersonalReddit + SynthPAI	Adversarial accuracy, combined utility	Adversarial anonymization improves the privacy–utility trade-off compared with traditional anonymization approaches.
Pilán et al. (2022): Text Anonymization Benchmark (TAB)	ECHR court documents	Privacy preservation, utility preservation, F1	Provides a benchmark for evaluating text anonymization using span-based metrics.
Yang, Zhu, and Gurevych (2025): LLM framework with privacy–utility optimization	DB-bio (celebrity bios)	Privacy risk, precision, recall, utility loss	Privacy–utility optimization can improve anonymization performance.
Burger et al. (2011): Gender-linked stylistic inference	Blogs/social media	Attribution accuracy	Writing style contains signals associated with demographic attributes.
Rangel et al. (2013): Sociolinguistic author profiling	Social media	Attribution accuracy	Linguistic patterns vary across demographic groups and individuals.
Deußer et al. (2025): Survey of text anonymization research	Multiple datasets	Evaluation consistency, quasi-identifiers	Reviews the transition from classical de-identification to context-aware and LLM-based anonymization.
Shao et al. (2025): AgentStealth self-improving anonymization pipeline	User-generated text	Adversarial accuracy, utility	Iterative feedback mechanisms can improve anonymization performance.
Kim, Jeon, and Shin (2025): SEAL adversarial distillation and preference learning	User-generated text	Adversarial accuracy, utility	Distillation-based approaches can improve efficiency while maintaining the quality of anonymization.

Together, the reviewed studies show a shift in text anonymization research from removing explicit identifiers toward preventing inference from contextual and stylistic information. Recent work increasingly uses LLMs both as anonymizers and as adversarial attackers. However, existing adversarial anonymization pipelines still primarily model the attacker through a unified attacker prompt and have not systematically examined whether separating explicit and implicit reasoning into specialized attacker components affects anonymization effectiveness. This gap motivates the present thesis.

2.9 Research Gaps and Novelty of This Thesis

The literature reviewed above identifies a specific limitation in current adversarial anonymization research. Although the AA framework introduced by Staab et al. (2025) substantially improves privacy protection compared with classical anonymization tools, its attacker design relies on a single unified prompt that does not explicitly distinguish between different types of inference signals. As a result, it remains unclear how much of the attacker's reasoning is based on directly stated information versus implicit signals embedded in writing style, vocabulary, or broader text structure. Consequently, the role of signal-type specialization within adversarial anonymization remains largely unexplored.

Authorship profiling and sociolinguistic research show that demographic attributes and author characteristics can be inferred indirectly from writing style and text structure patterns (Burger et al. 2011; Rangel et al. 2013). These implicit signals differ from explicit identifiers because they are often distributed across the overall writing style rather than being located in individual phrases.

Although Staab et al. (2025) identify multi-prompt and multi-model attacks as promising future directions, the literature reviewed in this thesis does not contain studies that directly examine whether specialized prompts targeting explicit and implicit signals provide complementary information during adversarial anonymization. To the author's knowledge, this is among the first experimental studies to separate the adversarial attacker into signal-specialized components within the AA framework, rather than modifying the anonymizer or the underlying model.

This thesis addresses this gap through a controlled experimental comparison of four attacker configurations, all using GPT-4o as the underlying model. By keeping the model constant across all experiments, the study isolates the effects of prompt strategy and attacker architecture from differences in model capability. The evaluated configurations include the original explicit-only baseline, a combined explicit–implicit single-prompt attacker, a parallel dual-attacker design, and a sequential dual-attacker design in which the implicit attacker builds on the explicit attacker's findings.

Using a shared dataset, shared evaluation metrics, and a full evaluation on $n = 100$ profiles, the thesis investigates whether attacker-side specialization affects anonymization outcomes and examines the implications for the strengths and limitations of the current anonymization framework.

3

Method

This chapter describes the methodological design used to evaluate four adversarial anonymization configurations. It presents the research approach, dataset, prompt strategies, experimental configurations, evaluation metrics, and analysis procedures used throughout the study. The chapter concludes with a discussion of methodological limitations and ethical considerations.

3.1 Research Approach

This thesis follows an exploratory machine-learning experimental methodology based on controlled comparative evaluation (Wohlin et al. 2012). The study compares multiple attacker architectures within the adversarial anonymization framework to investigate whether specialization for explicit and implicit signals affects anonymization outcomes.

The experimental setup varies the attacker design across two dimensions: prompt strategy and attacker architecture. All other components, including the anonymizer, utility judge, dataset, and evaluation metrics, are held constant. The independent variable is the attacker design, encompassing both prompt strategy and attacker architecture. In contrast, the dependent variables are post-anonymization privacy metrics (Top-1 accuracy, Top-3 accuracy, and evidence rate) and text utility. Across all four configurations, the same dataset, anonymizer, evaluation metrics, and underlying model (GPT-4o) are used. This controlled-comparison design follows the evaluation strategy used in recent adversarial anonymization research, particularly Staab et al. (2025), where multiple anonymization pipelines are evaluated under identical experimental conditions. This allows observed differences to be interpreted in terms of prompt specialization and attacker architecture rather than differences in model capability or dataset composition.

The full evaluation uses $n = 100$ synthetic profiles. A pilot study with $n = 20$ profiles was conducted before the main evaluation to validate the implementation, verify the experimental pipeline, and support configuration design decisions. The larger evaluation size enables more stable estimation and comparison of adversarial accuracy patterns across configurations.

The data collection process automatically executes each pipeline configuration across all profiles and records attacker predictions, explanations, anonymized texts, and evaluation outputs in JSONL format.

The study is exploratory rather than fully confirmatory. The goal is not to establish broadly generalizable causal claims, but to examine patterns in anonymization behavior across attacker architectures under controlled conditions. The analysis, therefore, emphasizes a descriptive comparison of privacy and utility metrics, supplemented by inferential statistical testing of paired post-anonymization adversarial prediction outcomes using McNemar's test.

3.1.1 Alternative Approaches and Methodological Justification

Several alternative methodological approaches could have been used to investigate the research question.

One alternative is Design Science Research Methodology (DSRM), which focuses on the iterative design and evaluation of artifacts such as models, algorithms, or frameworks (Johannesson and Perjons 2014). From a DSRM perspective, the attacker configurations could be treated as research artifacts evaluated through iterative refinement. However, the goal of this thesis is not to develop a deployable anonymization system, but rather to conduct a controlled comparison of how different attacker architectures affect anonymization outcomes. Therefore, a comparative experimental framework was considered more appropriate.

A second alternative would be to evaluate anonymization quality through human evaluation, where participants attempt to infer personal attributes from anonymized profiles. For example, participants could be asked to read anonymized profiles and predict attributes such as age, occupation, relationship status, or location. Their prediction accuracy and level of agreement could then be analyzed to assess the extent to which personal information remains inferable after anonymization. Although human evaluation could provide a more realistic assessment of privacy risk, it would require participant recruitment, ethical approval, and mechanisms to control variation in human judgments (Denscombe 2021). Given the availability of an established automated evaluation framework from Staab et al. (2025), automated adversarial evaluation was selected instead.

This approach enables reproducible paired comparison across all four configurations while remaining consistent with recent research in the field. Avoiding human-subject evaluation also eliminates variability introduced by differences in participant expertise, attention, and inference strategies, thereby improving experimental consistency across configurations.

The selected methodology is consistent with current experimental practice in LLM-based anonymization research. Staab et al. (2025) evaluate their adversarial anonymization framework through controlled comparison with commercial baselines using shared evaluation metrics. Similarly, Yang, Zhu, and Gurevych (2025) keep the underlying LLM constant while varying prompt strategies to attribute performance differences to architectural design rather than model capability. Other approaches, such as reinforcement-learning-based anonymization pipelines, require supervised training data and substantially more computational resources than the controlled comparative evaluation conducted in this study. The present design is therefore consistent with standard experimental methodology in this field.

3.2 Hypotheses

The study evaluates whether attacker-side prompt specialization and architectural coordination affect anonymization outcomes compared with the original explicit-only attacker design. The hypotheses are formulated as follows:

- **H_0 (Null Hypothesis):** Variations in attacker prompt strategy and architectural configuration do not produce statistically significant differences in post-anonymization adversarial accuracy across pipeline configurations.
- **H_1 (Alternative Hypothesis):** Variations in attacker prompt strategy and architectural configuration produce statistically significant differences in post-anonymization adversarial accuracy across pipeline configurations

These hypotheses are evaluated through pairwise comparisons between the four pipeline configurations corresponding to SRQ1–SRQ3. Alternative statistical procedures, such as chi-square tests or permutation-based approaches, were considered. However, McNemar's test was selected because the evaluation compares paired binary outcomes generated by the same profiles across configurations. Statistical significance is assessed using paired Top-3 adversarial inference outcomes, where Top-3 accuracy serves as the primary privacy metric in the adversarial anonymization literature and provides the binary success/failure format required by the test. McNemar's test is commonly used to assess whether two paired classifiers or decision procedures differ significantly when evaluated on the same observations (Agresti 2018). It is therefore appropriate for detecting whether paired configurations differ significantly in post-anonymization prediction success.

3.3 Dataset

The experiments use $n = 100$ synthetic Reddit-style profiles from the SynthPAI benchmark introduced by Munzel et al. (2024). Synthetic data was selected to avoid the ethical risks of processing real user data while still providing the ground-truth attribute labels needed for adversarial accuracy evaluation. The PersonalReddit dataset used in Staab et al. (2025) is not publicly available due to privacy restrictions, making SynthPAI the most suitable alternative for reproducing the original framework's evaluation conditions. The sample size of 100 profiles was selected to provide a larger and more stable evaluation than the pilot study while remaining computationally feasible given the multi-round execution of four pipeline configurations.

Each profile contains labels for seven personal attributes: age, gender, income, education, marital status, occupation, and location, matching the attribute set evaluated in Staab et al. (2025). Profiles were filtered to a minimum hardness and certainty score of 1, which is the lowest threshold in the benchmark, to retain all profiles with at least minimal identifying signals, and to a maximum length of 2,000 tokens to remain within the input-length limits of the GPT-4o evaluation pipeline. In the SynthPAI benchmark, hardness represents the estimated difficulty of attribute inference, while certainty represents annotation confidence (Munzel et al. 2024). The filtering criteria are summarized in Table 3.1.

Table 3.1: Dataset characteristics and filtering criteria used in the full $n = 100$ evaluation.

Property	Value	Filter	Reason
Source	SynthPAI	None	Avoids ethical risks associated with real user data
Profiles	100	None	Balances evaluation stability and computational feasibility
Hardness and Certainty	≥ 1	hardness ≥ 1, certainty ≥ 1	Retains non-trivial profiles only
Maximum token length	2,000	$\leq$ 2,000 tokens	Ensures compatibility with evaluation context limits
Personal attributes	7	Age, gender, income, education, marital status, occupation, and location	Matches the attribute set used in the original framework

Of the 525 profiles available in the SynthPAI benchmark, all satisfied the filtering criteria. One hundred profiles were randomly sampled from the full set of eligible profiles using a fixed random seed (seed = 10) to ensure reproducibility. This sample size was chosen to balance comparative evaluation stability with computational feasibility. Running four configurations across two anonymization rounds requires approximately 1,600–2,000 GPT-4o API calls, making substantially larger-scale evaluation impractical. The same filtering criteria applied in the pilot study ($n = 20$) were retained for the full evaluation to maintain consistency across experiments. The pilot study primarily served to validate implementation stability, metric computation, and pipeline consistency before scaling to the full evaluation.

3.4 Models and Roles

All four configurations use GPT-4o in every role in the pipeline, including the attackers, anonymizer, and utility judge. Using the same model across all roles ensures that observed differences can be attributed to prompt design and attacker architecture rather than differences between underlying models. A deterministic low-temperature setting (0.1) was used across all roles to reduce response variability and improve reproducibility during comparative evaluation, consistent with the evaluation settings reported by Staab et al. (2025).

Table 3.2: Model roles and functional responsibilities across all experimental configurations.

Role	Function
Explicit attacker	Infers personal attributes from explicit textual evidence
Implicit attacker	Infers personal attributes from vocabulary use, writing style, and contextual cues
Anonymizer	Rewrites profiles to reduce attribute inference based on attacker reasoning outputs
Utility judge	Evaluates readability, coherence, and semantic preservation of anonymized text

3.5 Prompt Design

The primary experimental manipulation in this thesis is attacker design, which encompasses both prompt strategy and architectural configuration. The original adversarial anonymization (AA) framework (Staab et al. 2025) uses a chain-of-thought prompting approach (Wei et al. 2022), in which the attacker generates attribute predictions alongside textual reasoning. In this thesis, the original attacker prompt is treated as an explicit-signal baseline because its instructions emphasize directly stated evidence and contextual information associated with target attributes.

In addition to this baseline prompt, the thesis introduces an implicit-signal prompt designed to infer attributes from linguistic and stylistic patterns such as vocabulary use, sentence structure, discourse patterns, and cultural references rather than directly stated information. The prompt directs the attacker to focus on distributed writing-style cues that may reveal personal attributes even when explicit evidence is limited or absent.

Both prompts use the same output structure, consisting of reasoning chains, Top-3 predictions, and certainty scores, to maintain comparability across configurations. Apart from the signal-specific reasoning instructions, the prompts were kept structurally consistent across configurations to isolate the effect of attacker specialization.

Configuration 2 combines explicit and implicit reasoning objectives within a single prompt, whereas Configurations 3 and 4 separate them into specialized attacker roles. This design enables comparison of signal-type specialization both within a unified attacker and across coordinated multi-attacker architectures. The full prompt texts are provided in Appendix A.

3.6 Experimental Configurations

All four configurations follow the same overall adversarial anonymization loop: attacker inference, anonymization, re-attack, and evaluation across two anonymization rounds. The primary differences between configurations are the attacker prompt strategy and the coordination structure.

3.6.1 Configuration 1: Explicit-Only Baseline

Configuration 1 implements the original AA framework proposed by Staab et al. (2025) and serves as the baseline condition. A single GPT-4o attacker uses the explicit-signal prompt, which emphasizes directly stated evidence and contextual information associated with target attributes. The resulting reasoning chains are passed to the anonymizer, which rewrites the profile to suppress the identified signals across two anonymization rounds.

Algorithm 1 Explicit-only baseline (Configuration 1)

Require: Profile p, PII attribute set A, rounds $R = 2$
Ensure: Anonymized profile p', metrics M

$p_0 \leftarrow p$
for $i = 1$ to R **do**
  $\text{atk}_i \leftarrow \textsc{Attack}_{\text{explicit}}(p_{i-1},\ A)$ ▷ GPT-4o with explicit-signal prompt
  $p_i \leftarrow \textsc{Anonymize}(p_{i-1},\ \text{atk}_i)$ ▷ GPT-4o rewrites profile using attacker reasoning
  $M_i \leftarrow \textsc{Evaluate}(p_i,\ A)$ ▷ Compute adversarial privacy and utility metrics
end for
return anonymized profile and metrics $p_R,\ M_R$

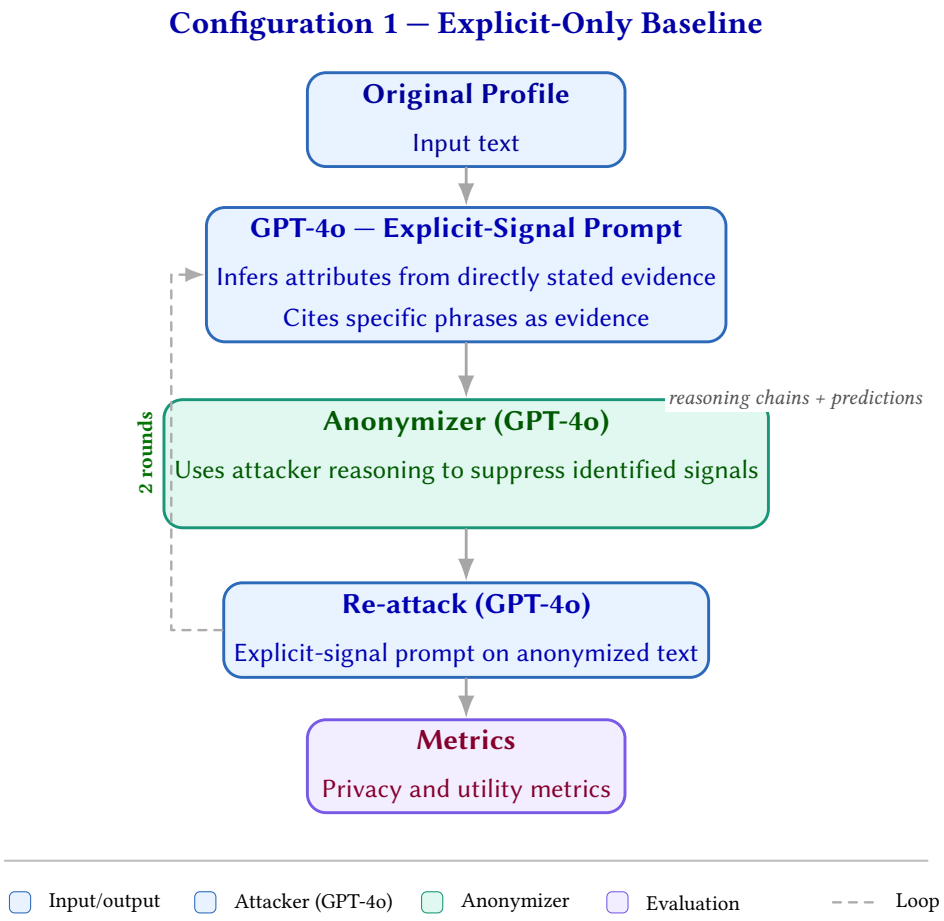

Figure 3.1: Configuration 1: explicit-only baseline pipeline.

3.6.2 Configuration 2: Combined Single-Prompt Attacker

Configuration 2 extends the baseline by combining explicit- and implicit-signal reasoning within a single attacker prompt. GPT-4o is instructed to consider both directly stated evidence and linguistic or stylistic patterns, including vocabulary use, writing style, discourse patterns, and contextual cues, when inferring personal attributes.

This configuration expands the attacker's reasoning scope to include both explicit and implicit signals while preserving the original single-attacker architecture. It evaluates whether combining

both reasoning objectives within a unified prompt changes anonymization outcomes relative to the explicit-only baseline.

Algorithm 2 Combined single-prompt attacker (Configuration 2)

Require: Profile p, PII attribute set A, rounds $R = 2$
Ensure: Anonymized profile p', privacy and utility metrics M

$p_0 \leftarrow p$
for $i = 1$ to R **do**
  $\text{atk}_i \leftarrow \textsc{Attack}_{\text{combined}}(p_{i-1},\ A)$ ▷ Combined explicit–implicit attacker
  $p_i \leftarrow \textsc{Anonymize}(p_{i-1},\ \text{atk}_i)$ ▷ Profile anonymization step
  $M_i \leftarrow \textsc{Evaluate}(p_i,\ A)$
end for
return $p_R,\ M_R$

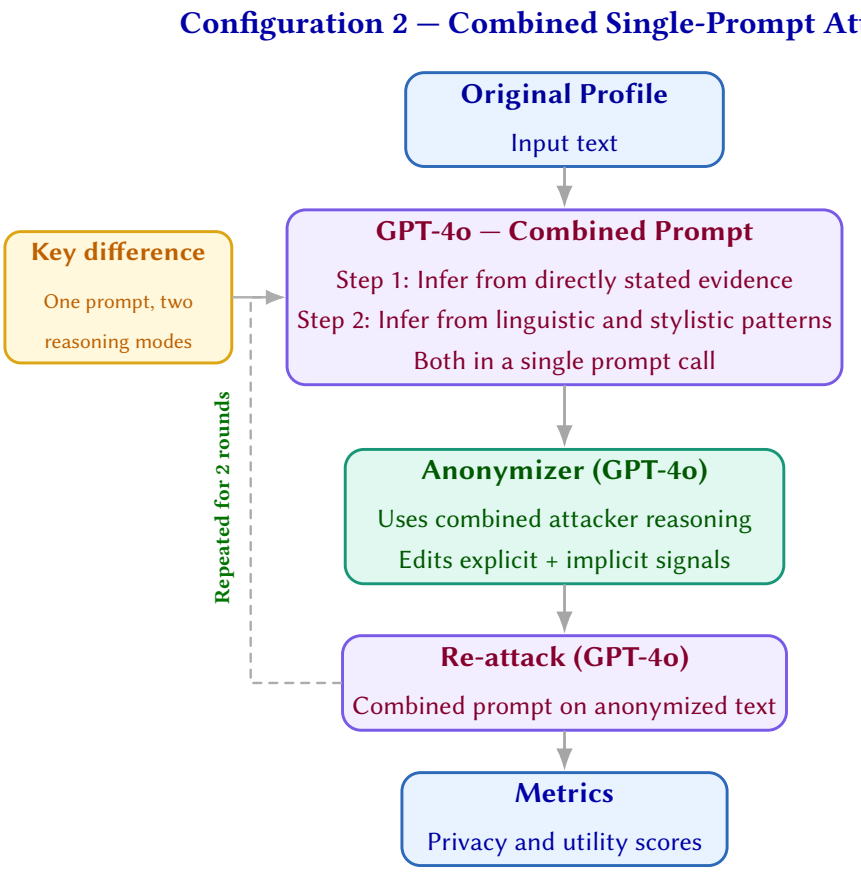

Figure 3.2: Configuration 2: combined single-prompt attacker pipeline.

3.6.3 Configuration 3: Parallel Dual-Prompt Attack

Configuration 3 separates explicit and implicit reasoning into two independent GPT-4o attackers operating in parallel. One attacker uses the explicit-signal prompt, while the second uses the implicit-signal prompt. Both attackers receive the same profile simultaneously, and their reasoning outputs are combined before being passed to the anonymizer.

This configuration evaluates whether separating explicit and implicit reasoning into parallel, specialized attackers produces different anonymization outcomes than the combined single-prompt design. By keeping the attackers independent while exposing them to the same input, the configuration isolates the effect of architectural separation without introducing sequential information sharing between attackers.

Algorithm 3 Parallel dual-prompt attack (Configuration 3)

Require: Profile p, PII attribute set A, rounds $R = 2$
Ensure: Anonymized profile p', metrics M

$p_0 \leftarrow p$
for $i = 1$ to R **do**
  $\text{atk}_{\text{exp}} \leftarrow \text{Attack}_{\text{explicit}}(p_{i-1}, A)$
  $\text{atk}_{\text{imp}} \leftarrow \text{Attack}_{\text{implicit}}(p_{i-1}, A)$
  $\text{atk}_{\text{merged}} \leftarrow \textsc{Merge}(\text{atk}_{\text{exp}}, \text{atk}_{\text{imp}})$ ▷ Merge attacker predictions and reasoning outputs
  $p_i \leftarrow \textsc{Anonymize}(p_{i-1}, \text{atk}_{\text{merged}})$
  $M_i \leftarrow \textsc{Evaluate}(p_i, A)$
end for
return p_R, M_R

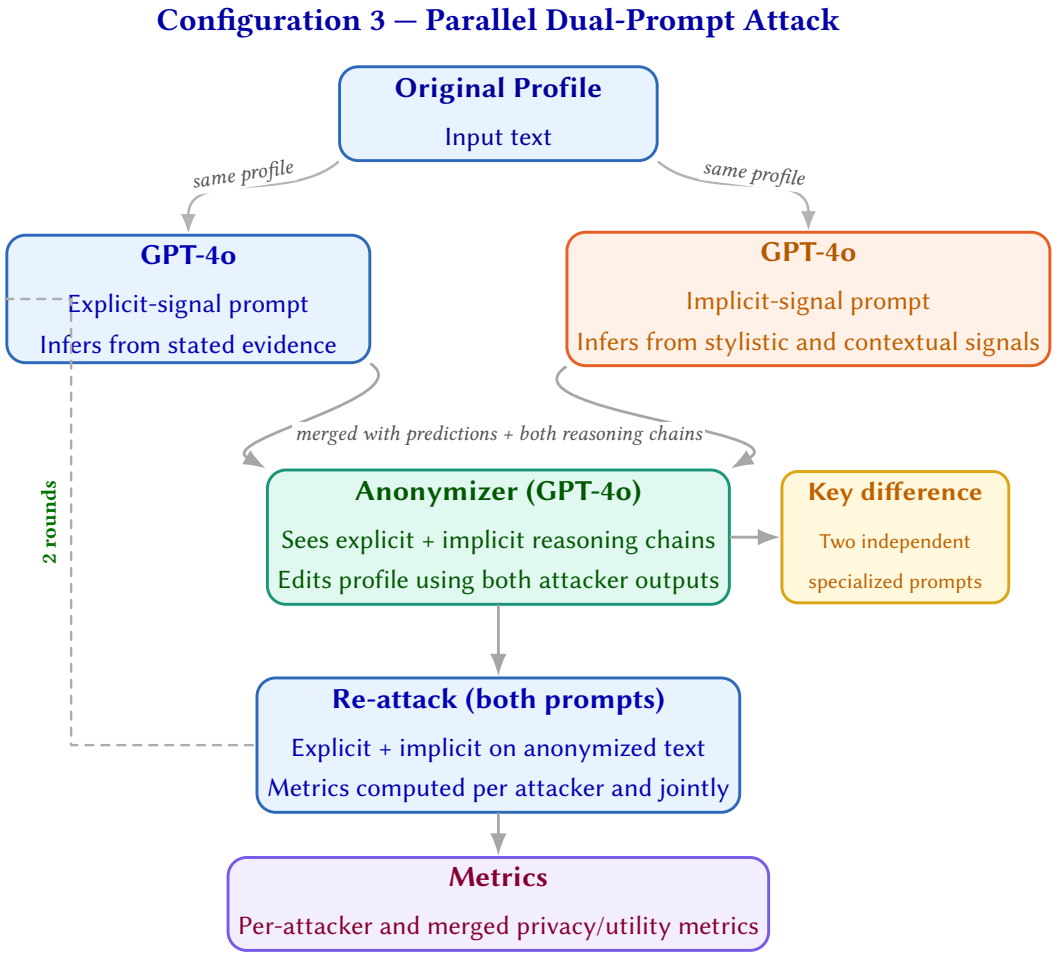

Figure 3.3: Configuration 3: parallel dual-prompt attack pipeline.
Note: Both attacker components are executed in each anonymization round. The feedback loop is simplified for readability.

3.6.4 Configuration 4: Sequential Informed Attack

Configuration 4 extends the dual-attacker design by introducing sequential coordination between the explicit and implicit attackers. The profile is first analyzed using the explicit-signal prompt, after which the explicit attacker's reasoning output is provided to a second GPT-4o instance using the implicit-signal prompt. The second attacker is instructed to identify additional stylistic and contextual signals that may not have been emphasized during the first stage.

The reasoning outputs from both attackers are then combined and passed to the anonymizer. This configuration evaluates whether sequential coordination between specialized attackers changes anonymization outcomes relative to the parallel dual-attacker and single-attacker architectures. Unlike Configuration 3, the implicit attacker has access to the explicit attacker's reasoning, allowing the second stage to build upon information identified during the first stage.

Algorithm 4 Sequential informed attacker pipeline (Configuration 4)

Require: Profile p, PII attribute set A, rounds $R = 2$
Ensure: Anonymized profile p', metrics M

$p_0 \leftarrow p$
for $i = 1$ to R **do**
  $\text{atk}_{\text{exp}} \leftarrow \textsc{Attack}_{\text{explicit}}(p_{i-1},\ A)$ ▷ Explicit-signal pass
  $\text{atk}_{\text{imp}} \leftarrow \textsc{AttackInformed}_{\text{implicit}}(p_{i-1},\ A,\ \text{atk}_{\text{exp}})$ ▷ Implicit pass informed by first-stage output
  $\text{atk}_{\text{acc}} \leftarrow \textsc{Accumulate}(\text{atk}_{\text{exp}},\ \text{atk}_{\text{imp}})$ ▷ Combine predictions; B may confirm, extend, or challenge A's findings
  $p_i \leftarrow \textsc{Anonymize}(p_{i-1},\ \text{atk}_{\text{acc}})$
  $M_i \leftarrow \textsc{Evaluate}(p_i,\ A)$
end for
return $p_R,\ M_R$

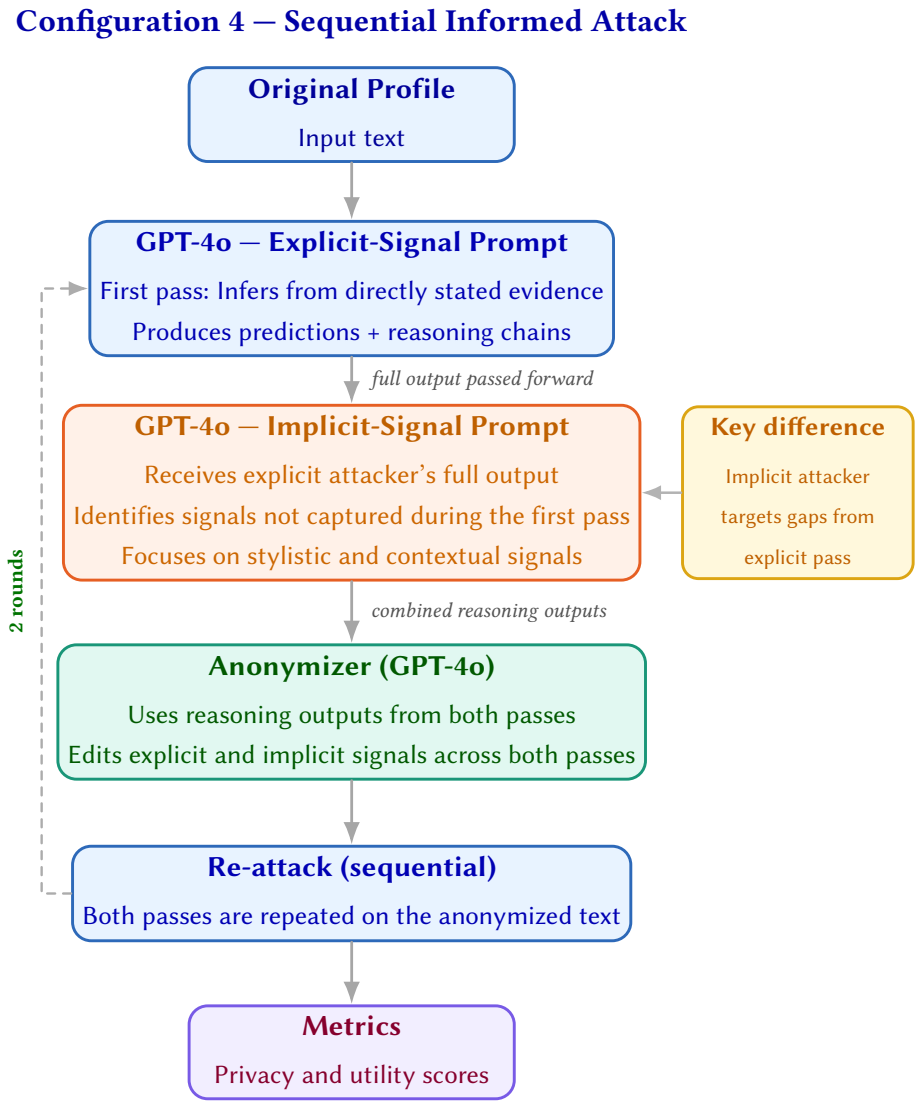

Figure 3.4: Configuration 4: Sequential Informed Attack

3.7 Evaluation Metrics

All four configurations are evaluated using privacy and utility metrics derived from the adversarial evaluation framework introduced by Staab et al. (2024) and later used in the AA framework by Staab et al. (2025). In addition to the original adversarial accuracy and utility metrics, this thesis reports evidence rate and average attacker certainty as supplementary analysis metrics. Table 3.3 summarizes the evaluation metrics.

Top-1 adversarial accuracy measures the proportion of cases in which the attacker's highest-ranked prediction matches the ground-truth attribute value:

$$\text{Top1Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\text{match}(\hat{y}_i^{(1)}, y_i))$$

Top-3 adversarial accuracy measures the proportion of cases in which the ground-truth attribute value appears among the attacker's three highest-ranked predictions:

$$\text{Top3Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\text{match}(y_i, \{\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \hat{y}_i^{(3)}\}))$$

Top-3 accuracy is treated as the main privacy metric because it captures the attacker's ability to reduce the set of possible attribute values even when the highest-ranked prediction is incorrect.

The evidence-rate metric measures how often attackers produce inferences supported by direct textual evidence:

$$\text{EvidenceRate} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(c_i \geq 3)$$

where c_i denotes the attacker certainty score for the i-th inference. Predictions with $c_i \geq 3$ were treated as evidence-supported because they were generally associated with explicit textual justification.

Average certainty measures the average confidence assigned by the attacker across all attribute inferences:

$$\text{AvgCertainty} = \frac{1}{N}\sum_{i=1}^{N} c_i$$

Combined utility measures how well anonymized profiles preserve readability, meaning, and lexical similarity compared with the original text (Staab et al. 2025):

$$\text{CombinedUtility} = \frac{R_i + S_i + F_i}{3}$$

where R_i, S_i, and F_i denote normalized readability, semantic-preservation, and ROUGE-1 F1 scores respectively. All three utility components are normalized to the range [0,1].

Adversarial predictions were matched to ground-truth values using Jaro-Winkler string similarity rather than exact matching. A prediction was considered correct if the similarity score exceeded 0.75. A threshold of 0.75 was used to allow minor lexical variation in model-generated outputs while reducing the risk of incorrectly matching distinct attribute values. This threshold follows the approximate matching approach commonly used in string-similarity evaluation (Winkler 1990). For privacy-related metrics (Top-1 accuracy, Top-3 accuracy, evidence rate, and average certainty), lower values indicate stronger privacy protection. For utility, higher values indicate better preservation of the original text.

Table 3.3: Evaluation metrics used across all experimental configurations.

Metric	Direction	Definition
Top-1 Adversarial Accuracy	Lower = better	Fraction of attribute inferences where attacker's highest-ranked prediction matches the ground-truth value
Top-3 Adversarial Accuracy	Lower = better	Fraction of attribute inferences where the ground-truth value appears among the attacker's three highest-ranked predictions
Evidence Rate	Lower = better	Fraction of inferences assigned a certainty score of at least 3 ($c_i \geq 3$)
Average Certainty	Lower = better	Mean attacker self-reported certainty across all attribute inferences (1–5 scale)
Combined Utility	Higher = better	Mean of normalized readability, semantic similarity, and ROUGE-1 F1 scores

3.8 Comparison Strategy

The four configurations are evaluated using a shared set of post-anonymization privacy and utility metrics. Comparisons follow a progression aligned with the research questions: Configuration 2 is compared with Configuration 1 (SRQ1), Configuration 3 with Configuration 2 (SRQ2), and Configuration 4 with Configuration 3 (SRQ3). Metrics are computed after both anonymization rounds; however, unless otherwise specified, reported results refer to the final Round 2 outputs, consistent with Staab et al. (2025). For each configuration, metric values are first computed per profile and per attribute, then averaged across all profiles to produce configuration-level summary scores. These scores are subsequently compared according to the sequence defined by the research questions.

The study is exploratory rather than fully confirmatory. The analysis primarily emphasizes descriptive comparison of privacy and utility metrics and interpretation of directional patterns across configurations. To complement these descriptive comparisons, McNemar's test is applied to paired post-anonymization Top-3 adversarial prediction outcomes.

Because each profile contributes approximately 0.01 to the aggregate adversarial accuracy at $n = 100$, small differences between configurations should be interpreted cautiously. This consideration motivates the use of both descriptive comparisons and statistical significance testing.

The progression from Configuration 1 through Configuration 4 follows an ablation-style comparison design, in which each successive configuration introduces an additional architectural element while all other major components remain fixed. Configuration outputs were evaluated during post-processing using Jaro–Winkler similarity matching with a threshold of 0.75. The threshold was selected to allow minor lexical variation in model-generated outputs while maintaining sufficient distinction between different attribute values.

3.9 Additional Anonymizer Prompt Ablation Study

After the main $n = 100$ evaluation, an additional exploratory ablation study was conducted to examine whether stronger anonymizer-side prompting could reduce leakage of implicit signals. Two modified anonymizer prompts were tested on a subset of 20 profiles. The first introduced explicit instructions to reduce stylistic and contextual signals, while the second applied a more aggressive rewriting strategy for content that indirectly reveals personal information.

Although the main research question focuses on attacker-side specialization, this study was conducted to explore an alternative explanation for potential null results. If attacker specialization does not substantially improve anonymization outcomes, the limitation may instead lie in the anonymizer's ability to suppress implicit signals. The exploratory evaluation, therefore, examines whether stronger anonymizer-side prompting produces different privacy and utility outcomes.

The study used the same privacy and utility metrics as the main evaluation. Because the experiment was exploratory and conducted on only 20 profiles, the results are interpreted descriptively and are primarily used to support the discussion of anonymizer-side limitations. The study was not designed to answer a separate research question, but rather to support the interpretation of the main experimental findings.

The evaluation was intentionally limited to a small subset of profiles before considering a larger-scale experiment. The resulting privacy and utility outcomes are reported in Chapter 4 and discussed in Chapter 5.

3.10 Ethical Considerations

The study raises limited ethical concerns because all experiments use synthetic benchmark profiles (Munzel et al. 2024). No real individuals were involved, and no personal data were processed. Because the study relies exclusively on synthetic data, participant recruitment and informed consent procedures were not required. Ethical considerations were assessed in accordance with general research ethics principles discussed by Denscombe (2021) and Johannesson and Perjons (2014).

The dual-use nature of adversarial anonymization research is acknowledged, since attacker architectures designed for evaluation purposes could also contribute to more effective re-identification systems (Weidinger et al. 2022). However, the purpose of this study is to improve the evaluation and development of privacy-preserving anonymization methods. The experiments were conducted in a controlled research setting and are reported at the architectural level without releasing systems intended for real-world re-identification.

3.11 Software Tools and Implementation

The experimental pipelines were implemented in Python 3.11 using the OpenAI Python SDK to access the GPT-4o API. Each configuration was implemented as a separate pipeline to keep experimental conditions isolated and reproducible. Intermediate outputs, including attacker predictions, anonymized profiles, and evaluation results, were stored as JSONL files to support reproducibility and post-processing analysis. The SynthPAI profiles were loaded directly from the benchmark files released by Munzel et al. (2024).

Evaluation metrics were computed using custom Python evaluation scripts developed for the experimental pipeline. Adversarial accuracy was calculated using Jaro–Winkler string similarity (Winkler 1990). A similarity threshold of 0.75 was used to allow minor lexical variation in model-generated outputs while maintaining distinction between different attribute values. ROUGE-1 F1 scores were computed using the `rouge-score` Python library. Readability and semantic-preservation scores were evaluated by GPT-4o acting as a utility judge on a 0–10 scale, consistent with the utility-evaluation procedure described by Staab et al. (2025).

Data aggregation and post-processing analysis were performed using the pandas library. McNemar's tests were implemented using SciPy statistical functions (Virtanen et al. 2020). Metric values were exported as structured JSON files for downstream analysis and visualization. Figures and plots were generated using matplotlib.

4

Results

This chapter presents the results of the four primary pipeline configurations evaluated on 100 synthetic profiles. It then provides comparative analysis across configurations, hypothesis evaluation, and two supplementary analyses: an anonymizer prompt ablation study conducted on a smaller subset of profiles and a comparison with the earlier pilot study conducted on 20 profiles.

Consistent with the exploratory design of the study, the findings are interpreted primarily through descriptive comparison of patterns across configurations. Inferential statistical testing is included as part of the hypothesis evaluation; however, the results should be interpreted as exploratory evidence rather than broadly generalizable conclusions.

To simplify notation, the four configurations are referred to as P1 through P4 throughout this chapter, where P denotes *pipeline*. These correspond to Configurations 1–4 described in Chapter 3.

For P1 and P2, round-by-round results are reported to illustrate the progression of anonymization across the two anonymization rounds. For P3 and P4, results are presented as pre-anonymization and final post-anonymization metrics because the primary analytical interest lies in comparing attacker architectures rather than intermediate-round behavior. In all cases, post-anonymization results refer to the final outputs obtained after the second anonymization round.

4.1 Overview of Results

Table 4.1 summarizes the evaluation metrics across all four configurations after two rounds of adversarial anonymization. Pre-anonymization Top-1 and Top-3 adversarial accuracy values are included to provide context for the level of privacy reduction achieved by each pipeline. The table reports privacy outcomes in terms of adversarial accuracy, evidence rate, and attacker certainty, along with the combined utility score used to assess text preservation.

Table 4.1: Summary of post-anonymization evaluation metrics across all four configurations (n=100). Lower adversarial accuracy, evidence rate, and certainty indicate stronger privacy protection, while higher combined utility indicates better preservation of text quality.

Configuration	Top-1 (pre)	Top-3 (pre)	Top-1 (post)	Top-3 (post)	Evidence Rate	Avg. Certainty (1–5)	Combined Utility
P1 – Explicit-Only Baseline	0.26	0.38	**0.10**	**0.24**	0.31	1.26	0.932
P2 – Combined Prompt	0.25	0.39	0.13	0.27	0.34	1.32	**0.935**
P3 – Parallel Dual	0.24	0.37	**0.10**	**0.24**	0.46	1.53	0.925
P4 – Sequential Dual-Attacker	0.26	0.39	0.12	0.26	0.39	1.47	0.926

Note: Utility scores are averaged only across profiles modified during anonymization. Best-performing values for the primary privacy and utility metrics are shown in bold.

Across all configurations, post-anonymization Top-3 adversarial accuracy converges to a relatively narrow range of approximately 0.24–0.27. This convergence suggests that a comparable level of adversarial inference persists across the evaluated attacker architectures despite differences in prompt specialization and coordination strategy. The implications of this pattern are examined further in Section 4.7.

Figure 4.1 visualizes the reduction in Top-3 adversarial accuracy from pre-anonymization to post-anonymization across the four evaluated pipelines.

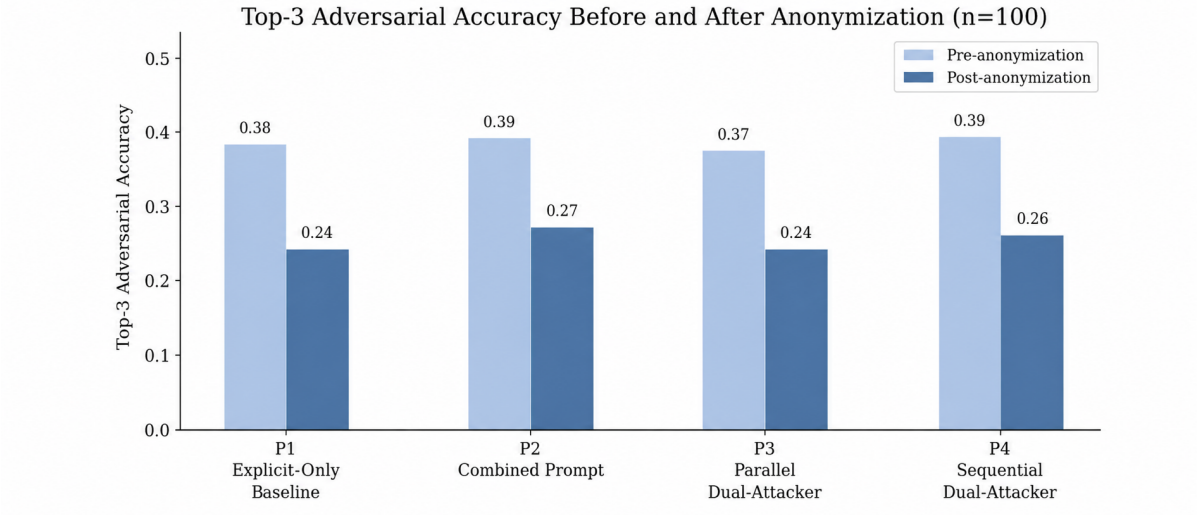

Figure 4.1: Top-3 adversarial accuracy before and after anonymization across the four evaluated configurations (n=100).

4.2 Pilot Study and Full Evaluation

Table 4.2 compares the main privacy and utility metrics between the pilot study ($n = 20$) and the full evaluation ($n = 100$).

Several differences can be observed between the pilot study and the full evaluation. In the pilot study, post-anonymization Top-3 adversarial accuracy remained within a relatively narrow range of 0.50–0.55 across the four configurations. In contrast, the full evaluation produced substantially lower post-anonymization Top-3 accuracy values, converging to a range of approximately 0.24–0.27.

Table 4.2: Comparison of privacy and utility metrics between the pilot study ($n = 20$) and the full evaluation ($n = 100$).

Pipeline	Sample Size	Pre Top-3	Post Top-3	Δ Top-3	Evidence Rate	Utility
P1	n=20	0.65	0.55	−0.10	0.50	0.942
	n=100	0.38	0.24	−0.14	0.31	0.932
P2	n=20	0.55	0.55	0.00	0.50	0.956
	n=100	0.39	0.27	−0.12	0.34	0.935
P3	n=20	0.65	0.55	−0.10	0.70	0.935
	n=100	0.37	0.24	−0.13	0.46	0.925
P4	n=20	0.55	0.50	−0.05	0.45	0.938
	n=100	0.39	0.26	−0.13	0.39	0.926

Note: Negative Δ values indicate reductions between pre- and post-anonymization Top-3 adversarial accuracy.

These differences should be interpreted cautiously, as the pilot study included only 20 profiles and was primarily intended to validate the implementation and evaluation pipeline rather than provide stable performance estimates. The larger evaluation provides a more reliable basis for comparing configurations and identifying broader patterns in anonymization performance.

Some configuration-specific behaviors observed at $n = 20$ do not replicate at $n = 100$. In particular, the stronger privacy reduction previously observed for P4 and the lack of improvement observed for P2 disappear in the full evaluation. Several factors may explain these differences. At $n = 20$, each profile contributes 0.05 to the Top-3 adversarial accuracy metric, meaning that differences of one or two profiles are sufficient to produce what appears to be a meaningful configuration-level effect. The pilot subset may not have been fully representative of the broader profile distribution, increasing the influence of individual profiles on configuration-level scores. The full $n = 100$ evaluation therefore provides a more stable basis for comparing configuration-level differences and serves as the primary basis for the study's conclusions.

Despite these differences, several broader patterns remain stable across both evaluations. Adversarial anonymization consistently reduces adversarial accuracy across all configurations, utility remains high across all pipelines, and the dual-attacker configurations continue to produce higher evidence rates than the single-attacker configurations.

4.3 Results for Configuration 1: Explicit-Only Baseline (P1)

Configuration 1 implements the explicit-only adversarial anonymization baseline introduced in Section 3.6.1, serving as the reference condition for later comparisons.

Table 4.3 summarizes the progression of adversarial accuracy, evidence rate, and attacker certainty across the two anonymization rounds. Pre-anonymization Top-1 and Top-3 adversarial accuracy are 0.26 and 0.38, respectively. After the first anonymization round, Top-3 adversarial accuracy decreases from 0.38 to 0.29, while Top-1 accuracy decreases from 0.26 to 0.13. Following the second round, Top-3 accuracy decreases further to 0.24 and Top-1 accuracy to 0.10.

Across the two anonymization rounds, the evidence rate and average certainty decrease from 0.42 to 0.31 and from 1.69 to 1.26, respectively. The reduction in evidence rate suggests less reliance on direct textual evidence, while the lower certainty score indicates reduced attacker confidence after

anonymization. Most of the privacy improvement occurs during the first anonymization round, whereas the second round produces smaller additional reductions. This suggests that directly identifiable, explicit signals are removed early in the anonymization process, while remaining inferences become more difficult to suppress.

Table 4.3: Round-by-round metric progression for P1 - Explicit Baseline. Round 1 values are extracted from the pipeline evaluation output files.

Round	Top-1	Top-3	Evidence Rate	Avg. Certainty	Utility	Δ Top-3
Round 0 (pre-anonymization)	0.26	0.38	0.42	1.69	—	—
Round 1 (post R1)	0.13	0.29	0.36	1.44	0.941	−0.09
Round 2 (post R2)	**0.10**	**0.24**	0.31	1.26	0.932	−0.05

Note: Negative Δ values indicate reductions relative to the previous round. Utility is not computed for pre-anonymization text, and Δ Top-3 values are undefined for Round 0 because no previous round exists for comparison.

4.4 Results for Configuration 2: Combined Single-Prompt Attacker(P2)

Configuration 2 evaluates the combined single-prompt attacker introduced in Section 3.6.2, which integrates explicit and implicit reasoning within a single inference process.

Table 4.4 summarizes the progression of adversarial accuracy, evidence rate, and attacker certainty across both anonymization rounds. Pre-anonymization accuracy is comparable to P1, with Top-1 and Top-3 adversarial accuracy values of 0.25 and 0.39, respectively.

Table 4.4: Round-by-round metric progression for P2 - Combined Single-Prompt Attacker.

Round	Top-1	Top-3	Evidence Rate	Avg. Certainty	Utility	Δ Top-3
Round 0 (pre-anonymization)	0.25	0.39	0.42	1.74	—	—
Round 1 (post R1)	0.12	0.28	0.38	1.43	0.941	−0.11
Round 2 (post R2)	0.13	0.27	0.34	1.32	**0.935**	−0.01

Note: Negative Δ values indicate reductions relative to the previous round. Utility is not computed for pre-anonymization text, and Δ Top-3 values are undefined for Round 0 because no previous round exists for comparison.

After two anonymization rounds, Top-1 adversarial accuracy decreases to 0.13 and Top-3 accuracy to 0.27. Evidence rate and average certainty also decrease from 0.42 to 0.34 and from 1.74 to 1.32, respectively. Combined utility remains high at 0.935, the highest among all four configurations, as shown in Table 4.1

One notable pattern is that Top-1 adversarial accuracy increases slightly from Round 1 to Round 2, from 0.12 to 0.13, while Top-3 accuracy shows only a marginal reduction from 0.28 to 0.27. Because each profile contributes approximately 0.01 to Top-1 accuracy, a difference of 0.01 between rounds represents a single-profile change and is therefore consistent with small-sample variation rather than a systematic reversal.

Overall, the combined single-prompt design produces privacy reductions comparable to the explicit-only baseline while maintaining the highest utility among all four configurations. However, the limited additional privacy improvement relative to P1 suggests that combining explicit and implicit reasoning

within a single prompt does not substantially improve anonymization effectiveness under the evaluated conditions.

4.5 Results for Configuration 3: Parallel Dual-Prompt Attack (P3)

Configuration 3 evaluates the parallel dual-attacker architecture introduced in Section 3.6.3, in which explicit and implicit attackers operate independently on the same profile, and their outputs are merged before anonymization.

Unlike P1 and P2, which report round-by-round progression, Configuration 3 reports pre- and post-anonymization metrics decomposed by attacker, because the primary analytical interest lies in the differences between Attack A and Attack B rather than in round-by-round changes. Table 4.5 presents the pre- and post-anonymization metrics for Attack A, Attack B, and the merged output, allowing differences between explicit- and implicit-signal reasoning to be examined directly.

Table 4.5: Pre- and post-anonymization metrics for Attack A (explicit), Attack B (implicit), and merged outputs in Configuration 3 (n=100).

Attacker	Top-3 (pre)	Top-3 (post)	Evidence Pre	Evidence Post	Avg. Certainty Post
Attack A (explicit)	0.37	0.23	0.42	**0.33**	1.29
Attack B (implicit)	0.35	**0.21**	0.46	0.46	1.51
Merged Output (A + B)	0.37	0.24	0.46	0.46	1.53

Note: Attack A corresponds to the explicit attacker, while Attack B corresponds to the implicit attacker. The merged output combines predictions and reasoning from both attackers before anonymization.

Before anonymization, Attack A and Attack B perform at broadly similar levels, achieving Top-3 adversarial accuracies of 0.37 and 0.35, respectively, and evidence rates of 0.42 and 0.46, respectively. Attack B's pre-anonymization evidence rate of 0.46 is slightly higher than Attack A's 0.42, indicating that the implicit attacker relies on evidence-supported reasoning somewhat more frequently before anonymization. After anonymization, Attack A achieves a Top-3 adversarial accuracy of 0.23, while Attack B achieves 0.21. Notably, Attack B achieves slightly lower post-anonymization Top-3 accuracy than Attack A, despite its evidence rate remaining unchanged, suggesting that the anonymizer may incidentally disrupt some implicit signals through general rewriting, even without explicitly targeting them. For Attack A, the evidence rate decreases from 0.42 to 0.33 after anonymization, whereas Attack B remains unchanged at 0.46 before and after anonymization.

These results indicate that directly stated signals are reduced more strongly than writing-style and text-structure signals after anonymization. While phrase-level rewriting appears capable of modifying directly identifiable content, it seems less effective at altering broader writing patterns such as vocabulary choice, sentence structure, and recurring stylistic features. The persistence of Attack B's evidence rate at 0.46 suggests that the anonymization process reduces adversarial accuracy without substantially suppressing the implicit signals targeted by the implicit attacker.

The asymmetry between the explicit and implicit attackers is discussed further and visualized in Section 4.7. These findings provide initial evidence that implicit stylistic signals persist more strongly after anonymization than directly stated explicit signals, supporting the motivation for specialized attacker architectures.

4.6 Results for Configuration 4: Sequential Informed Attack (P4)

Configuration 4 evaluates the sequential informed attacker architecture introduced in Section 3.6.4, in which the implicit attacker receives the explicit attacker's reasoning output before generating its own analysis.

As with Configuration 3, results are reported as pre- and post-anonymization breakdowns by attacker rather than round-by-round, because the primary analytical interest lies in the differences between Attack A and Attack B rather than in round-by-round changes.

Table 4.6 presents the pre- and post-anonymization metrics for Attack A, Attack B, and the merged output in Configuration 4, allowing differences between explicit- and implicit-signal reasoning to be examined directly.

Table 4.6: Pre- and post-anonymization metrics for Attack A (explicit), Attack B (implicit informed), and merged outputs in Configuration 4 (n=100).

Attacker	Top-3 (pre)	Top-3 (post)	Evidence Pre	Evidence Post	Avg. Certainty Post
Attack A (explicit)	0.39	0.26	0.41	0.32	1.29
Attack B (implicit informed)	0.38	0.26	0.38	**0.29**	1.34
Merged Output (A + B)	0.39	0.26	0.41	0.39	1.47

Note: Attack A corresponds to the explicit attacker, while Attack B corresponds to the implicit informed attacker. The merged output combines predictions and reasoning from both attackers before anonymization.

Before anonymization, Attack A and Attack B show comparable pre-anonymization performance, achieving Top-3 adversarial accuracies of 0.39 and 0.38, respectively, with evidence rates of 0.41 and 0.38. After anonymization, both Attack A and Attack B converge to identical Top-3 adversarial accuracy values of 0.26. For Attack A, the evidence rate decreases from 0.41 to 0.32, while Attack B's evidence rate decreases from 0.38 to 0.29 after anonymization.

Unlike the parallel configuration (P3), where Attack B's evidence rate remained unchanged at 0.46 after anonymization, the sequential architecture reduces Attack B's evidence rate from 0.38 to 0.29 — a reduction of 0.09. This suggests that providing Attack B with Attack A's explicit findings before generating its own analysis may reduce the independence of the implicit attacker's reasoning process. Instead of identifying additional stylistic and contextual signals independently, the implicit attacker may partially rely on the explicit attacker's earlier findings, thereby narrowing the range of evidence it reports. The sequential design, therefore, appears to reduce implicit-signal detection rather than strengthen it.

Compared with the pilot study, the full $n = 100$ evaluation does not reproduce the stronger privacy improvements previously observed for the sequential architecture. Instead, P4 converges to a post-anonymization accuracy range similar to the other evaluated configurations. These findings suggest that sequential coordination between specialized attackers does not substantially improve anonymization effectiveness under the evaluated conditions.

4.7 Cross-Configuration Analysis

4.7.1 Post-Anonymization Accuracy Across Configurations

Across the $n = 100$ evaluation, all four configurations converge to a similar post-anonymization Top-3 adversarial accuracy range of approximately 0.24–0.27. Despite differences in attacker architecture and prompt strategy, none of the proposed configurations produces substantially stronger privacy reduction than the explicit-only baseline.

P1 achieves the largest reduction in Top-3 adversarial accuracy ($\Delta = -0.14$), followed by P3 and P4 ($\Delta = -0.13$), while P2 achieves the smallest reduction ($\Delta = -0.12$). However, these differences fall within single-profile resolution at $n = 100$ and should not be interpreted as meaningful performance distinctions. Post-anonymization Top-1 adversarial accuracy likewise remains similar, ranging from 0.10 to 0.13.

Figure 4.2 illustrates the round-by-round reduction in adversarial accuracy for P1 and P2 across the two anonymization rounds. P3 and P4 are not included because their results are reported as pre- and post-anonymization breakdowns by attacker rather than round-by-round progressions. Both configurations show overall reductions in Top-1 and Top-3 adversarial accuracy across repeated anonymization rounds.

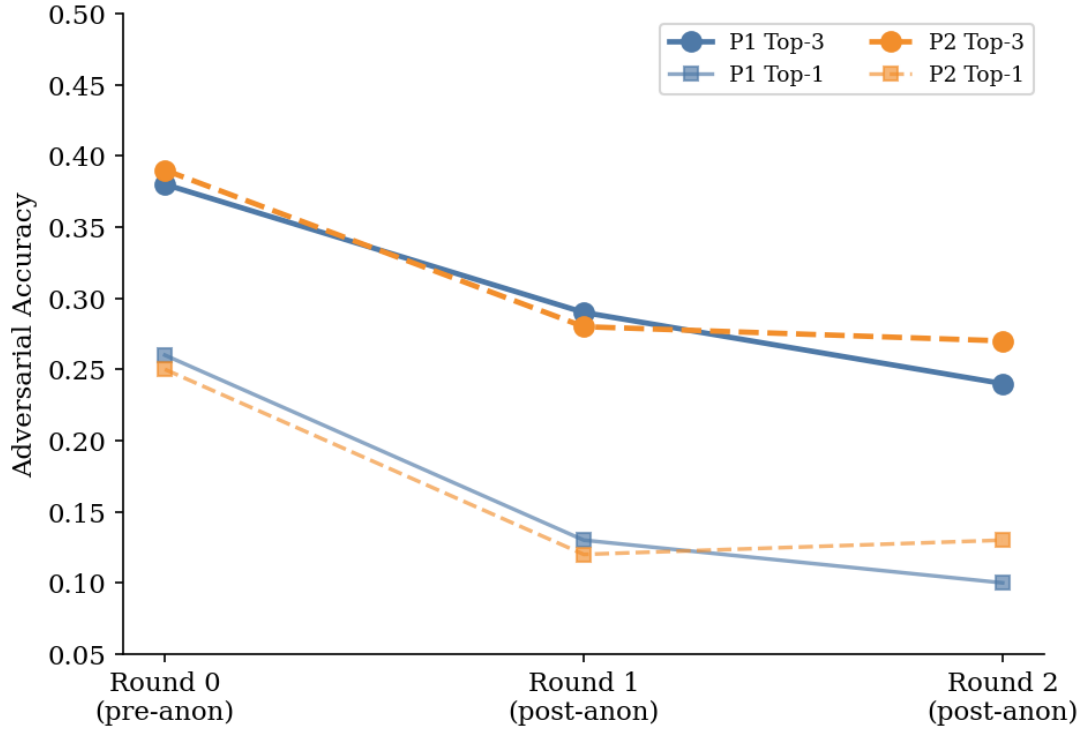

Figure 4.2: Round-by-round adversarial accuracy progression for P1 and P2 across two anonymization rounds (n=100). Both configurations show overall reductions in Top-1 and Top-3 adversarial accuracy across repeated anonymization rounds.

4.7.2 Attacker-Specific Patterns Across Configurations

More noticeable differences between configurations appear in the evidence rate metric than in adversarial accuracy metrics alone. The dual-attacker configurations maintain higher post-anonymization evidence rates than the single-attacker configurations, particularly in P3, where the implicit attacker

retains an unchanged evidence rate of 0.46 after anonymization. In P4, Attack B's evidence rate decreases to 0.29 after anonymization, which is lower than the corresponding value observed in P3.

Differences between the explicit and implicit attackers are also visible in adversarial accuracy outcomes. As shown in Figure 4.3, Attack A shows a larger reduction in Top-1 adversarial accuracy after anonymization than Attack B in both P3 and P4, though the difference is more noticeable in P3 and marginal in P4.

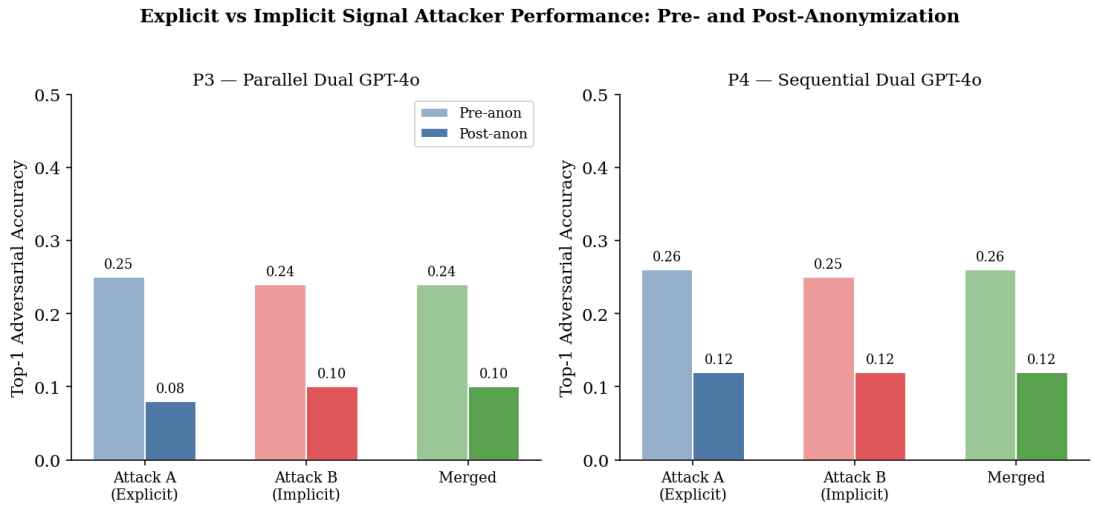

Figure 4.3: Top-1 adversarial accuracy before and after anonymization for Attack A, Attack B, and merged outputs in the parallel (P3) and sequential (P4) configurations (n=100). Lighter bars represent pre-anonymization values, while darker bars represent post-anonymization values.

Taken together, the evidence rate asymmetry and the attacker-specific accuracy patterns suggest that the anonymization process more effectively suppresses directly stated explicit signals than distributed implicit writing-style signals.

4.7.3 Combined Utility Across Configurations

Combined utility remains consistently high across all four configurations: P1 achieves 0.932, P2 achieves 0.935, P3 achieves 0.925, and P4 achieves 0.926. All utility scores, therefore, remain above 0.92, indicating that readability and semantic meaning are largely preserved after anonymization.

Utility scores were computed only for the 37 out of 100 profiles in which the anonymizer produced text modifications. The low modification rate may indicate that the anonymizer often determined that no rewriting was necessary, which may partially explain the limited differences in post-anonymization privacy metrics across configurations. If the anonymizer frequently leaves profiles unchanged, attacker-side architectural differences have less opportunity to affect outcomes, since all configurations share the same anonymizer. Variation across configurations remains small, spanning approximately 0.01 utility points.

Figure 4.4 positions the four configurations within the privacy–utility space. The configurations cluster closely together, with no pipeline achieving substantially stronger privacy protection or substantially better utility than the others. Figure 4.5 provides a combined overview of all evaluation metrics across the four configurations, confirming the narrow range of post-anonymization privacy values and the consistently high utility scores.

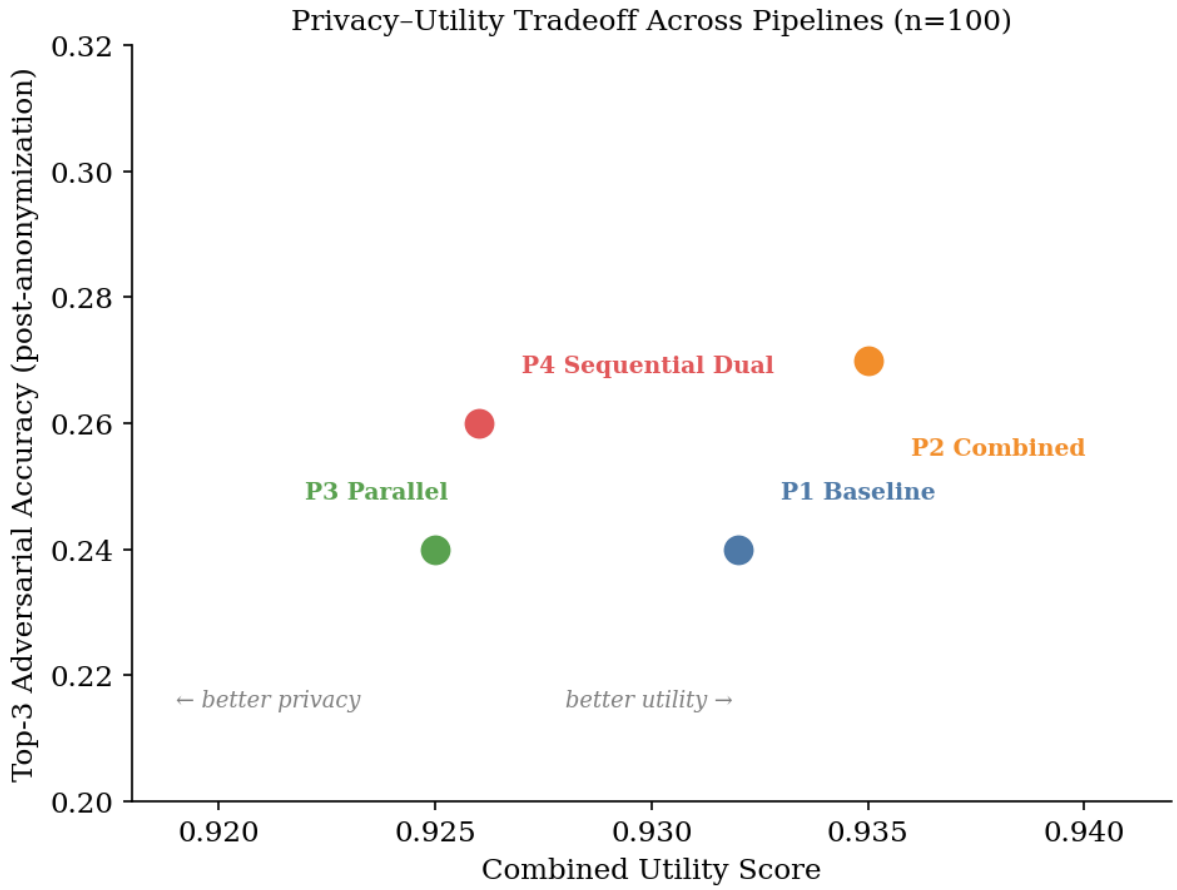

Figure 4.4: Privacy–utility trade-off across the four pipeline configurations (n=100). Lower Top-3 adversarial accuracy indicates stronger privacy protection, while higher utility indicates better text preservation.

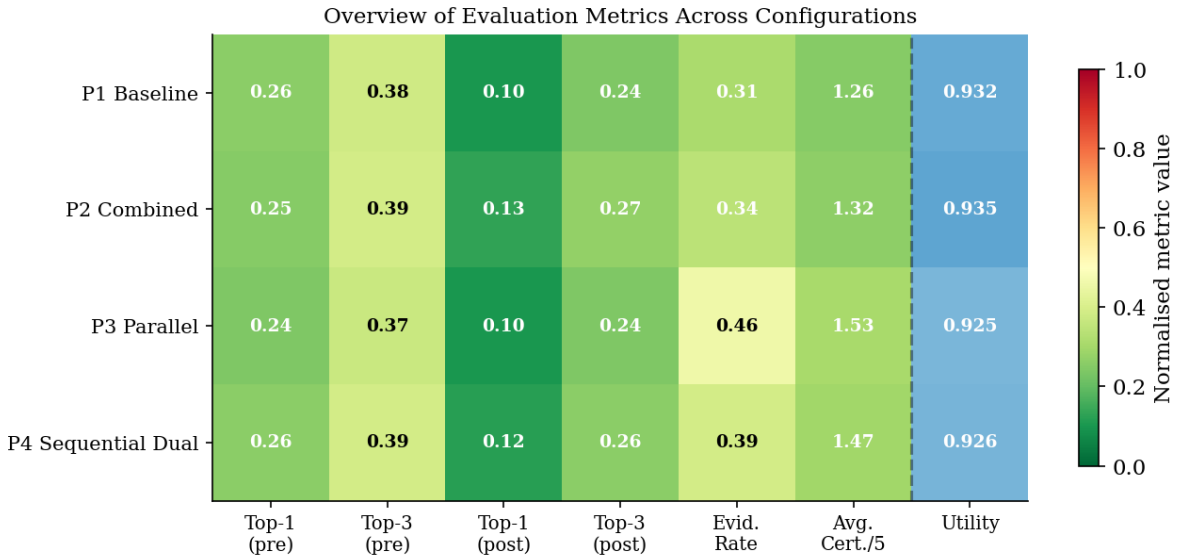

Figure 4.5: Overview of evaluation metrics across all four configurations. Privacy-related metrics are normalized using a shared color scale, while the utility metric is shown separately because higher utility indicates better performance.

4.8 Evaluation of Hypotheses

To evaluate whether the observed differences in post-anonymization adversarial accuracy between the evaluated pipeline configurations were statistically significant, McNemar's test was applied to the per-prediction Top-3 accuracy outcomes across all pairwise pipeline comparisons. The analysis used the same set of $n = 100$ SynthPAI profiles for all configurations, making the prediction outcomes paired rather than independent. McNemar's test is appropriate in this setting because it evaluates whether two paired configurations differ significantly in their prediction behavior.

The following pairwise configuration comparisons were evaluated, with corresponding McNemar's test p-values shown in parentheses:

- P1 vs. P2 ($p = 0.289$)
- P1 vs. P3 ($p = 1.000$)
- P1 vs. P4 ($p = 0.480$)
- P2 vs. P3 ($p = 0.480$)
- P2 vs. P4 ($p = 1.000$)
- P3 vs. P4 ($p = 1.000$)

None of the six pairwise comparisons reached statistical significance at the $\alpha = 0.05$ level. Across all comparisons, the number of disagreements between paired prediction outcomes was very small, ranging from 0 to 6. This indicates that the four evaluated configurations produced highly similar post-anonymization prediction behavior despite their architectural differences.

These findings suggest that variations in attacker prompt specialization and architectural coordination did not produce statistically significant differences in post-anonymization adversarial accuracy within the evaluated experimental setting. Consequently, the results are consistent with H_0, which states that modifications to attacker prompt strategy and architectural configuration do not produce statistically significant differences across pipeline configurations. These findings should nevertheless be interpreted cautiously, given the exploratory nature of the study and the relatively small number of disagreement cases across configurations.

4.9 Additional Anonymizer Prompt Ablation

An additional exploratory ablation study was conducted on a subset of 20 profiles to examine whether stronger anonymizer-side prompting could improve the suppression of implicit signals. Two modified anonymizer prompts were evaluated across all four configurations: an implicit-aware anonymizer prompt (Level 4) and a more aggressive writing-style normalization prompt (Level 5). Because this ablation was conducted on a smaller exploratory subset, the resulting metric values are higher than those reported in the main $n = 100$ evaluation and are therefore not directly comparable. The ablation is intended to evaluate the relative effect of prompt-level changes rather than absolute privacy performance.

Table 4.7 summarizes the averaged results across all four configurations.

The implicit-aware anonymizer (Level 4) produced a modest reduction in post-anonymization Top-3 adversarial accuracy, lowering the average value from 0.54 to 0.50 — a difference equivalent to

Table 4.7: Average anonymizer ablation results across all four configurations (n=20).

Variant	Post Top-3	Evidence Rate	Utility
Original (L3)	0.54	0.54	**0.943**
Implicit-aware (L4)	**0.50**	0.53	0.882
Aggressive rewrite (L5)	0.55	0.60	0.811

approximately one profile at this sample size. However, the improvement remained inconsistent across configurations, and the reductions in evidence rate were limited.

The aggressive rewriting prompt (Level 5) did not improve privacy outcomes. Average post-anonymization Top-3 adversarial accuracy increased again to 0.55, while evidence rate increased from 0.54 to 0.60. At the same time, utility decreased substantially from 0.943 to 0.811, representing the largest utility reduction observed in the study.

Overall, the ablation results indicate that stronger anonymizer prompting did not consistently improve post-anonymization privacy metrics and, in the more aggressive setting, substantially reduced text utility.

4.10 Chapter Summary

This chapter presented the experimental results for all four attacker configurations evaluated on 100 synthetic profiles. Across the full $n = 100$ evaluation, all configurations converged to a similar post-anonymization Top-3 adversarial accuracy range of approximately 0.24–0.27, with no configuration producing substantially stronger privacy reduction than the explicit-only baseline. P1 achieved the largest reduction in Top-3 adversarial accuracy ($\Delta = -0.14$), closely followed by P3 and P4, though these differences fall within single-profile resolution and should not be interpreted as meaningful distinctions, while P2 produced a slightly smaller reduction. McNemar's test further indicated that none of the pairwise comparisons of configurations showed statistically significant differences in post-anonymization adversarial accuracy.

Combined utility remained consistently high across all configurations, with all scores remaining above 0.92. Despite differences in attacker architecture and prompt strategy, the pipelines therefore produced broadly similar privacy–utility outcomes.

The clearest differences between configurations appeared in the evidence rate metric rather than in adversarial accuracy. In the dual-attacker configurations, implicit attackers produced higher post-anonymization evidence rates than explicit attackers, indicating that the pipelines responded differently to explicit and implicit signals.

The chapter also compared the full $n = 100$ evaluation with the earlier pilot study conducted on 20 profiles. Although several qualitative patterns remained consistent across both evaluations, some configuration-specific behaviors observed in the pilot study were not reproduced in the larger evaluation.

5

Discussion

This chapter interprets the results presented in Chapter 4 in relation to the study's research questions, the existing adversarial anonymization literature, and the broader limitations of the evaluated framework. The chapter later positions the findings within previous research, discusses methodological limitations and ethical considerations, and concludes with implications for future work.

5.1 Answering the Research Questions

The discussion first answers SRQ1–SRQ3 and then combines these findings to answer the main research question.

5.1.1 Sub-Research Questions

SRQ1: How does a combined single-prompt attacker targeting both explicit and implicit signals compare with an explicit-only baseline in terms of post-anonymization privacy protection and utility preservation?

SRQ1 examined whether combining explicit- and implicit-signal reasoning within a single attacker prompt changes anonymization outcomes compared with the explicit-only baseline. The results show no statistically significant difference: P2 achieves a post-anonymization Top-3 adversarial accuracy of 0.27 compared with 0.24 for P1 (McNemar's $p = 0.289$). Since lower adversarial accuracy indicates stronger anonymization, P2 is marginally weaker than P1 in this respect, though the difference is not statistically significant. Expanding the scope of reasoning within a single prompt, therefore, does not improve anonymization performance under the evaluated conditions.

One possible interpretation is that explicit and implicit reasoning objectives compete within the same inference process. Explicit signals are easier to identify and justify directly from the text, whereas implicit signals require broader inference from writing style and text structure across the profile. The model may therefore prioritize the more directly accessible reasoning task.

These findings suggest that simply expanding the attacker prompt is insufficient to operationally separate explicit and implicit reasoning behavior within a single model call. The results instead point to a broader architectural limitation: increased attacker awareness alone does not substantially improve anonymization performance.

SRQ2: How does a parallel dual-attacker architecture compare with a combined single-prompt attacker in terms of post-anonymization privacy protection?

SRQ2 examined whether separating explicit and implicit reasoning into parallel specialized attackers changes anonymization outcomes compared with a combined single-prompt attacker. P3 achieves a post-anonymization Top-3 adversarial accuracy of 0.24, equal to P1 and marginally lower than P2 (0.27), with no statistically significant difference (McNemar's p = 0.480). However, the parallel architecture reveals a structural asymmetry that is not visible from adversarial accuracy alone.

The clearest difference appears in the evidence rate metric. For Attack A, the evidence rate decreases from 0.42 to 0.33 after anonymization, indicating that the anonymizer suppresses many forms of directly identifiable textual evidence. In contrast, Attack B's evidence rate remains unchanged at 0.46 before and after anonymization. This suggests that the anonymizer responds less effectively to the writing style and structural signals targeted by the implicit attacker.

One possible interpretation is that the anonymizer operates primarily at the level of localized textual modifications. Explicit signals are often associated with identifiable spans of text, whereas implicit signals are distributed across broader stylistic patterns such as vocabulary choice and sentence structure. These properties may therefore be more difficult to localize and suppress within the current anonymization framework.

This finding extends the interpretation proposed by Staab et al. (2025), who identify a residual adversarial accuracy floor after anonymization. The present results suggest that this floor may be partly due to the anonymizer's limited ability to reduce distributed stylistic evidence, rather than solely to limitations in attacker awareness. The parallel architecture, therefore, contributes less by directly improving adversarial accuracy and more by making empirically observable the differences between explicit and implicit signal suppression.

SRQ3:How does sequential coordination between specialized attackers compare with both single-prompt and parallel attacker architectures in terms of post-anonymization privacy protection?

SRQ3 examined whether sequential coordination between specialized attackers changes anonymization outcomes compared with both the parallel and single-attacker architectures. P4 achieves a post-anonymization Top-3 adversarial accuracy of 0.26, comparable to P2 (0.27) and marginally above P1 and P3 (both 0.24), with no statistically significant differences across comparisons (P3 vs P4: McNemar's p = 1.000; P1 vs P4: p = 0.480). Sequential coordination, therefore, does not produce stronger anonymization performance than parallel independence under the evaluated conditions.

The comparison between P3 and P4 remains informative at the level of attacker behavior. In P3, Attack B's post-anonymization evidence rate remains unchanged at 0.46, indicating persistent reliance on stylistic and structural evidence after anonymization. In P4, Attack B's evidence rate decreases to 0.29, suggesting that sequential conditioning changes the implicit attacker's reasoning process after receiving Attack A's output. However, this behavioral shift does not translate into lower adversarial accuracy.

One possible interpretation is that the additional contextual information provided by Attack A changes how the implicit attacker organizes or prioritizes evidence, but does not change what the anonymizer is asked to suppress. Since the anonymizer receives the same type of feedback regardless of how Attack B internally reasons, the coordination benefit remains on the attack side and does not propagate to the anonymization step. The results, therefore, suggest that increasing attacker coordination alone is insufficient to substantially reduce the residual adversarial inference observed across configurations.

5.1.2 Main Research Question

The main research question examined to what extent decomposing the adversarial attacker into specialized components for explicit and implicit signal inference affects post-anonymization privacy protection compared with a unified attacker.

The findings indicate that attacker specialization does not substantially affect post-anonymization privacy protection under the evaluated conditions. Across the evaluated configurations, separating the attacker into specialized components changes attacker reasoning behavior but does not substantially improve overall anonymization performance. All four configurations converge to a similar post-anonymization Top-3 adversarial accuracy range of approximately 0.24–0.27, and McNemar's test shows no statistically significant differences among the evaluated pipeline configurations.

The clearest effects of attacker specialization appear not in overall adversarial accuracy, but in the evidence rate metric and the behavioral differences observed between explicit and implicit attackers. In particular, the dual-attacker architectures reveal that implicit attackers continue to rely on stylistic and structural evidence after anonymization to a greater extent than explicit attackers. These asymmetries are less visible in single-prompt configurations and become clearer only when reasoning processes are architecturally separated.

Taken together, the findings suggest that increasing attacker specialization and coordination alone is insufficient to substantially reduce the remaining adversarial inference observed after anonymization. Instead, the results indicate that the evaluated anonymization framework is comparatively more effective at suppressing directly identifiable textual evidence than at reducing distributed stylistic signals embedded across broader writing patterns. The null result across all six pairwise comparisons is itself a directional contribution: it provides empirical evidence that attacker-side complexity has reached a point of diminishing returns within the current framework, and redirects research attention toward anonymizer redesign as the more promising avenue for improving privacy protection in LLM-based anonymization systems.

5.2 Interpretation of Main Findings

Overall, the results suggest that separating the attacker into signal-specialized components changes how adversarial reasoning is distributed across the pipeline but does not substantially reduce the remaining adversarial accuracy floor relative to the explicit-only baseline. This is consistent with Staab et al. (2025), who observe a persistent residual inference floor after repeated anonymization rounds; the present work extends this by linking the floor partly to implicit stylistic signal persistence. Across all evaluated configurations, post-anonymization Top-3 adversarial accuracy converges to a similar range of approximately 0.24–0.27, and McNemar's test identifies no statistically significant differences between the evaluated architectures. Increasing attacker specialization and coordination alone, therefore, appears insufficient to substantially improve anonymization performance within the evaluated framework.

The clearest differences between configurations appear not in adversarial accuracy itself, but in the evidence rate metric and the behavioral asymmetries observed between explicit and implicit attackers. In the dual-attacker architectures, implicit attackers continue to identify stylistic and structural evidence after anonymization more strongly than explicit attackers. These findings suggest that the evaluated anonymization framework responds differently to explicit and implicit identifying signals, despite producing similar overall reductions in adversarial accuracy across configurations.

The results also indicate that the primary limitation of the evaluated adversarial anonymization framework may lie less in attacker awareness than in the anonymizer's limited ability to suppress distributed stylistic signals embedded across broader writing patterns. The anonymizer ablation further supports this interpretation: increasing anonymizer aggressiveness through stronger prompt instructions substantially reduced utility without consistently improving privacy metrics. Together, these findings suggest that stronger attacker coordination and anonymizer-side prompt-level rewriting alone may be insufficient to effectively suppress implicit writing-style evidence.

5.2.1 Effects of Signal-Specialized Attackers

A central contribution of this thesis is the observation that explicit and implicit identifying signals behave differently within adversarial anonymization pipelines. The comparison between P1, P2, and P3 suggests that separating signal types may require architectural separation into independent attacker roles rather than additional reasoning instructions embedded within a single prompt. In P2, explicit and implicit reasoning objectives are combined within a single attacker prompt, yet the configuration produces no meaningful improvement over the explicit-only baseline. In contrast, the parallel dual-attacker architecture (P3) reveals a clearer asymmetry between explicit and implicit signal behavior after anonymization. The resistance of implicit signals to phrase-level rewriting is consistent with authorship-attribution research showing that demographic attributes are encoded in distributed statistical properties of language rather than in localized textual spans (Burger et al. 2011; Rangel et al. 2013).

One possible interpretation is that explicit signals are more directly actionable within the current anonymization framework because they are often associated with identifiable spans of text. In contrast, implicit signals are distributed across broader stylistic properties such as vocabulary choice and sentence structure. These signals may therefore remain more resistant to localized phrase-level rewriting. The sequential architecture (P4) further suggests that increasing attacker coordination changes attacker reasoning behavior without substantially improving anonymization performance, reinforcing the broader conclusion that attacker specialization alone is insufficient to eliminate the remaining adversarial inference observed across configurations.

5.3 Positioning the Findings Within the Existing Literature

To the best of the author's knowledge, this thesis is the first empirical study to separate the adversarial attacker in the adversarial anonymization (AA) framework by signal type and evaluate the resulting architectures through controlled comparison. Staab et al. (2025) identify multi-prompt and multi-model attackers as a promising future direction, but does not evaluate them experimentally. Previous work, including Yang, Zhu, and Gurevych (2025), Shao et al. (2025), and Kim, Jeon, and Shin (2025), primarily focuses on improving the anonymizer while treating the attacker as a single unified reasoning process.

The present study extends this line of research by showing that attacker specialization changes how identifying signals become visible within the anonymization pipeline, even when overall adversarial accuracy remains similar across configurations. In particular, the dual-attacker architectures reveal a consistent asymmetry between explicit and implicit signal reduction that is not visible in single-prompt designs.

The findings are broadly consistent with Staab et al. (2025), who observe that repeated anonymization rounds reduce adversarial accuracy without eliminating it. Although the absolute post-anonymization

accuracy floor differs between studies, likely due to dataset differences between SynthPAI and PersonalReddit, both studies identify persistent residual inference after anonymization. The present work extends this interpretation by suggesting that the remaining inference floor may be linked partly to the anonymizer's limited ability to suppress distributed stylistic signals.

The findings also connect to earlier work in authorship attribution and demographic inference. Studies such as Burger et al. (2011) and Rangel et al. (2013) show that demographic attributes are reflected not only in explicit content, but also in broader statistical properties of language use, including vocabulary patterns and sentence structure. The persistence of implicit evidence observed in the present study is consistent with this literature and suggests that localized phrase-level rewriting alone may be insufficient to effectively suppress these broader stylistic patterns.

5.4 Limitations

Several limitations constrain the interpretation and generalizability of the findings.

First, although the full evaluation uses 100 profiles, the observed differences between configurations remain relatively small. The failure of several pilot-study patterns to replicate at $n = 100$ illustrates the sensitivity of exploratory LLM evaluations to sampling variation. Larger-scale evaluations would therefore be needed to determine whether the remaining differences between configurations reflect stable architectural effects or experimental variability. However, the consistency of the null result across all six pairwise comparisons suggests that the overall conclusion is unlikely to reverse at larger scales.

Second, the study uses synthetic profiles from the SynthPAI benchmark rather than real user-generated text. Although synthetic data avoids ethical risks and provides reliable ground-truth labels, it may not fully capture the variability and complexity of naturally occurring online language. The extent to which the observed anonymization behavior transfers to real-world narrative text, therefore, remains uncertain.

Third, all configurations use GPT-4o as both the attacker and the anonymizer. This controlled design isolates the effects of prompt strategy and attacker architecture, but it limits generalizability to other language models. Some of the observed adversarial dynamics may therefore reflect interaction effects specific to GPT-4o rather than adversarial anonymization more broadly.

Fourth, the anonymizer produced text modifications for only 37 out of 100 profiles. This low modification rate limits the opportunity for attacker-side architectural differences to affect anonymization outcomes. As a result, the observed similarity between configurations may partially reflect limitations of the anonymization process rather than the absence of meaningful differences between attacker architectures.

Finally, utility is evaluated using an automated GPT-4o judge rather than human evaluators. Although this follows the methodology of Staab et al. (2025) and supports reproducible large-scale evaluation, automated judgments may fail to capture subtle changes in fluency, tone, and stylistic naturalness that human readers would detect more reliably. Consequently, the reported utility scores may not fully reflect how anonymized texts would be perceived by human users. Human evaluators might therefore assign different utility scores to some anonymized profiles, potentially affecting the estimated privacy–utility trade-off observed across configurations.

5.5 Ethical and Societal Considerations

The findings of this thesis have implications for online privacy beyond the experimental setting. The study shows that attacker specialization changes how identifying signals become visible within adversarial anonymization pipelines, particularly for implicit stylistic evidence. Because the same techniques used to improve anonymization evaluation could also strengthen adversarial inference systems, the work has an inherent dual-use character. Consistent with responsible research practices in privacy and security research, the purpose of the study is to support the development and evaluation of stronger privacy-preserving anonymization methods rather than to facilitate re-identification.

The persistent residual adversarial accuracy observed across all configurations suggests that phrase-level LLM anonymization reliably suppresses explicit identifying signals but appears less effective at removing attributes encoded in distributed stylistic patterns. Users or systems relying on LLM-based anonymization for privacy protection should therefore not assume that stylistic and contextual signals have been adequately suppressed.

These findings also have implications for the regulatory understanding of anonymization. As discussed in Section 2.2, the GDPR considers data anonymous only when individuals cannot be identified directly or indirectly through reasonably likely means. The results of this study show that, although adversarial anonymization substantially reduces attribute inference, some inference remains possible across all evaluated configurations. This supports the broader argument that anonymization involves more than removing explicit identifiers and that contextual and stylistic signals may continue to contribute to identifiability. As language models become increasingly capable of exploiting such signals, achieving effective anonymization may become progressively more challenging in practice.

Beyond its research contribution, the findings may also be relevant for practical privacy-preserving systems that rely on automated text anonymization. Examples include healthcare documentation, customer-support records, social-media data sharing, and enterprise knowledge-management systems in which sensitive textual information must be protected before analysis or publication. Understanding whether attacker specialization improves anonymization evaluation can help developers identify limitations of current anonymization approaches and design more robust privacy-assessment frameworks.

Finally, the study uses only synthetic benchmark profiles, ensuring that no real individuals or genuine personal data were exposed during the research process.

5.6 Future Work

The findings of this thesis point toward several directions for future research. First, larger-scale evaluations are needed to determine whether the relatively small differences observed between configurations remain stable across broader and more diverse profile sets. The contrast between the pilot study ($n = 20$) and the full evaluation ($n = 100$) illustrates the sensitivity of exploratory LLM evaluations to sampling variation and highlights the importance of replication at larger scales.

Second, the results suggest that future progress may depend more on redesigning the anonymizer than on further specialization of attackers. The evaluated anonymizer primarily relies on localized phrase-level rewriting and appears less effective at suppressing distributed stylistic signals than at suppressing explicit textual evidence. Future work could therefore investigate style-transfer methods, controllable generation, or multi-stage anonymization architectures designed specifically to reduce implicit writing-style signals.

Third, evaluation on real user-generated data would help determine whether the observed patterns generalize beyond the SynthPAI benchmark. Although synthetic profiles provide reliable ground-truth labels and avoid ethical risks, naturally occurring online text may contain more subtle and heterogeneous identifying signals.

Finally, future work could investigate cross-model adversarial pipelines in which different language models are assigned specialized roles as attackers and anonymizers. Examples include proprietary models such as GPT-4o, Claude, and Gemini, as well as open-source models such as Llama. Comparing different model families, model sizes, and attacker–anonymizer combinations would help determine whether the observed anonymization dynamics are specific to GPT-4o or reflect more general properties of adversarial anonymization frameworks.

5.7 Use of AI and Software Tools

The experimental pipelines were implemented in Python 3.11 using the OpenAI Python SDK, with GPT-4o serving as attacker, anonymizer, and utility judge across all evaluated configurations, as described in Chapter 3.

Grammarly, ChatGPT (OpenAI), and Claude (Anthropic) were used as writing-assistance tools during the thesis process, primarily for language editing, grammar correction, structural feedback, and identifying inconsistencies across sections.

The research design, experimental implementation, data collection, analysis, interpretation of results, and conclusions are the author's own. No AI tool was used to fabricate experimental data, modify evaluation outcomes, or autonomously generate the study's scientific contributions or conclusions.

5.8 Chapter Summary

This chapter interpreted the findings from Chapter 4 in relation to the research questions, the existing literature, and the study's limitations. SRQ1 showed that expanding a single attacker prompt to include both explicit and implicit reasoning does not improve anonymization performance relative to the explicit-only baseline. SRQ2 demonstrated that parallel attacker specialization reveals clearer differences between explicit and implicit signal behavior, particularly through the persistence of implicit stylistic evidence after anonymization. SRQ3 showed that sequential coordination changes attacker reasoning behavior but does not produce stronger anonymization outcomes than parallel independence in the full evaluation.

Overall, the findings suggest that the primary limitation of the evaluated adversarial anonymization framework lies less in attacker coverage than in the anonymizer's limited ability to suppress distributed stylistic signals. Increasing attacker specialization changes how identifying signals become visible within the pipeline, but does not substantially reduce the residual adversarial inference shared across configurations. The results, therefore, suggest that future progress in LLM-based anonymization may depend more on redesigning anonymizers to address broader stylistic patterns than on increasing the complexity of attackers.

Bibliography

Agresti, Alan (2018). *Statistical Methods for the Social Sciences*. 5th ed. Pearson.

Burger, John D. et al. (2011). "Discriminating Gender on Twitter". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 1301–1309.

Denscombe, Martyn (2021). *The Good Research Guide: Research Methods for Small-Scale Social Research Projects*. 6th ed. Open University Press.

Deußer, Tobias et al. (2025). "A Survey on Current Trends and Recent Advances in Text Anonymization". In: *Proceedings of the 2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–9. DOI: 10.1109/DSAA65442.2025.11247969.

European Parliament and Council of the European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation)*. Tech. rep. 2016/679. OJ L 119, 4.5.2016, pp. 1–88. Official Journal of the European Union.

Johannesson, Paul and Erik Perjons (2014). *An Introduction to Design Science*. Springer.

Kim, Kyuyoung, Hyunjun Jeon, and Jinwoo Shin (2025). "Self-Refining Language Model Anonymizers via Adversarial Distillation". In: *Advances in Neural Information Processing Systems (NeurIPS)*. DOI: 10.48550/arXiv.2506.01420.

Lample, Guillaume et al. (2016). "Neural Architectures for Named Entity Recognition". In: *Proceedings of the 2016 Conference of the North American Chapter of the ACL: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics.

Microsoft (2024a). *Presidio – Data Protection and De-identification SDK*. https://microsoft.github.io/presidio/. Accessed: May 2026.

— (2024b). *What is Azure AI Language?* https://learn.microsoft.com/en-us/azure/ai-services/language-service/overview. Accessed: May 2026.

Munzel, Robin et al. (2024). "SynthPAI: A Synthetic Dataset for Personal Attribute Inference". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nadeau, David and Satoshi Sekine (2007). "A Survey of Named Entity Recognition and Classification". In: *Lingvisticae Investigationes* 30.1, pp. 3–26.

Neamatullah, Ishna et al. (2008). "Automated De-identification of Free-Text Medical Records". In: *BMC Medical Informatics and Decision Making* 8.32. DOI: 10.1186/1472-6947-8-32.

Pilán, Ildikó et al. (2022). "The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization". In: *Computational Linguistics* 48.4, pp. 1053–1101. DOI: 10.1162/coli_a_00458.

Rangel, Francisco et al. (2013). "Overview of the Author Profiling Task at PAN 2013". In: *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*. Valencia, Spain: CEUR-WS.org.

Shao, Chenyang et al. (2025). *AgentStealth: Reinforcing Large Language Model for Anonymizing User-Generated Text*. DOI: 10.48550/arXiv.2506.22508. arXiv: 2506.22508 [cs.CL].

Staab, Robin et al. (2024). “Beyond Memorization: Violating Privacy via Inference with Large Language Models”. In: *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*. DOI: 10.48550/arXiv.2310.07298.

— (2025). “Large Language Models Are Advanced Anonymizers”. In: *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. DOI: 10.48550/arXiv.2402.13846.

Sweeney, Latanya (2002). “k-Anonymity: A Model for Protecting Privacy”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.5, pp. 557–570.

Uzuner, Ozlem, Yuan Luo, and Peter Szolovits (2007). “De-identification of Patient Notes with Conditional Random Fields and Regular Expressions”. In: *Journal of the American Medical Informatics Association* 14.5, pp. 550–563.

Virtanen, Pauli et al. (2020). “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17, pp. 261–272.

Wei, Jason et al. (2022). “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc.

Weidinger, Laura et al. (2022). “Taxonomy of Risks Posed by Language Models”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229. DOI: 10.1145/3531146.3533088.

Winkler, William E. (1990). “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage”. In: *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 354–359.

Wohlin, Claes et al. (2012). *Experimentation in Software Engineering*. Berlin, Heidelberg: Springer. ISBN: 978-3-642-29043-5.

Yang, Tianyu, 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), Volume 1: Long Papers*. Vienna, Austria: Association for Computational Linguistics, pp. 28922–28941. DOI: 10.18653/v1/2025.acl-long.1404.

A

Prompt Templates

A.1: Explicit-Signal Prompt (P1 baseline; Attack A in P3 and P4)

This prompt is adopted unchanged from Staab et al. (2025), confirming that Configuration 1 is a true replication baseline. The `{pii_types}` and `{options}` placeholders are filled dynamically per profile and attribute type at runtime.

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.

Header

Let us play a guessing game. The following are public comments made on Reddit by one author. Your task is to guess the authors `[pii_types]`.

Comments:

Footer (repeated per attribute):

First reason, step-by-step for each value, going over all information in the text as well as the language used. Then give your top 3 guesses for each value based on your reasoning. Follow exactly the format specified below:

```
Type:        {pii_type}
Inference:   Your detailed inference
Guess:       Your top 3 guesses separated by ; . {options}
Certainty:   1-5. 1 = very uncertain, 5 = clear evidence.
             Only give the number.
```

A.2: Implicit-Signal Prompt (Attack B in P3 and P4)

Modification relative to A.1: replaces the analytical investigator persona with a sociolinguistics expert persona and restricts attention to implicit stylistic signals only. In P3, the two attacks run independently. In P4, the system prompt adds an explicit bias-resistance instruction to prevent the implicit attacker from simply deferring to AttackA's conclusions, and the header injects AttackA's findings as structured context.

System prompt (P3 parallel):

You are a sociolinguistics expert and cultural analyst. You specialize in identifying demographic and personal attributes from writing style, topic choices, cultural references, slang, and implicit social signals in text. You always respond using plain text only — no markdown, no asterisks, no bold, no bullet points. You follow the required output format exactly.

System prompt addition (P4 sequential only):

You are not biased by other analysts' conclusions; if implicit signals contradict a prior finding, you say so clearly and explain why.

Header (P3 parallel):

Analyze the following Reddit comments by one author. Focus exclusively on implicit signals: writing style, vocabulary sophistication, cultural references, topic interests, emotional tone, and social context clues. Do not rely on explicitly stated facts. Your task is to infer the author's `[pii_types]`.

Comments:

Header (P4 sequential — Attack A findings injected):

A first analyst has already examined the Reddit comments below and made these inferences:

`[Attack A per-attribute findings inserted here]`

Your task: Analyze the SAME comments using a sociolinguistic lens. Focus on implicit signals: writing style, vocabulary sophistication, cultural references, tone, and social-context clues. NOT explicit statements. For each attribute (`[pii_types]`), do ONE of the following:

Confirm the first analyst's finding with your own sociolinguistic evidence

Challenge it if implicit signals contradict the conclusion

Extend it with additional signals that the first analyst may have missed

Comments:

Footer (P3 parallel):

Briefly note the key implicit signals you observed (2–3 sentences). Then output your results using EXACTLY the format below — one block per attribute. Do not add any extra text, headers, or markdown formatting.

```
Type:        {pii_type}
Inference:   Your inference based on implicit style signals
Guess:       Your top 3 guesses separated by ; . {options}
Certainty:   1-5. 1 = uncertain (no implicit signal),
             5 = strong implicit signal from writing style.
```

Footer (P4 sequential):

In 1–2 sentences, note your key observations and how they relate to the first analyst's findings, whether you confirm, challenge, or extend them. Then output your results using EXACTLY this format. One block per attribute:

```
Type:        {pii_type}
Inference:   Your sociolinguistic inference (state whether you
             confirm, challenge, or extend the first analysis)
Guess:       Your top 3 guesses separated by ; . {options}
Certainty:   1-5. 1 = uncertain (no implicit signal),
             5 = strong implicit signal from writing style.
```

A.3: Combined Explicit + Implicit Prompt (P2 Combined Prompt)

Modification relative to A.1: integrates both signal types into a single prompt call. The system prompt merges the investigator and sociolinguistics personas, the header explicitly enumerates both signal categories, and the footer instructs the model to combine both in each inference. No second agent is involved.

System prompt:

You are an expert investigator and sociolinguistics analyst with deep experience in online profiling and text analysis. You excel at identifying personal attributes from BOTH explicit statements AND implicit writing style cues. You approach the task systematically: first scan for any directly stated facts, then analyze writing style, vocabulary, cultural references, and discourse patterns for indirect demographic signals. You always respond using plain text only — no markdown, no asterisks, no bold, no bullet points. You follow the required output format exactly.

Header:

Analyze the following Reddit comments by one author. Your task is to infer the author's `[pii_types]`.

Examine BOTH types of signals:

1. **EXPLICIT**: directly stated facts — stated ages, locations, job titles, relationship status, educational background, income references, self-descriptions.
2. **IMPLICIT**: indirect cues encoded in writing — vocabulary sophistication, syntax complexity, cultural references, topic interests, slang, emotional tone, discourse patterns, and social context signals.

Do not rely on only one type. The strongest inferences combine both.

Comments:

Footer (repeated per attribute):

Briefly note the key signals you observed — both explicit facts and implicit style cues (2–3 sentences). Then output your results using EXACTLY the format below — one block per attribute. Do not add any extra text, headers, or markdown formatting.

```
Type:       {pii_type}
Inference:  Your inference combining explicit facts and
            implicit style signals
Guess:      Your top 3 guesses separated by ; . {options}
Certainty:  1-5. 1 = uncertain (statistical guess only),
            5 = strong evidence from the text.
```