LARGE LANGUAGE MODELS ARE ANONYMIZERS

Robin Staab, Mark Vero, Mislav Balunović, Martin Vechev
Department of Computer Science
ETH Zurich {robin.staab,mark.vero}@inf.ethz.ch

ABSTRACT

Recent work in privacy research on large language models has shown that they achieve near human-level performance at inferring personal data from real-world online texts. With consistently increasing model capabilities, currently existing text anonymization methods are lacking behind both regulatory requirements and adversarial threats. This raises the question of how individuals can effectively protect their personal data in sharing online texts. In this work, we take two steps to answer this question: We first present a new setting for evaluating anonymizations in the face of adversarial LLMs inferences, allowing for a natural measurement of anonymization performance while remedying some of the shortcomings of previous metrics. We then present our LLM-based adversarial anonymization framework leveraging the strong inferential capabilities of LLMs to inform our anonymization procedure. In our experimental evaluation, we show on both real-world and synthetic online texts how adversarial anonymization outperforms current industry-grade anonymizers both in terms of the resulting utlity and privacy.

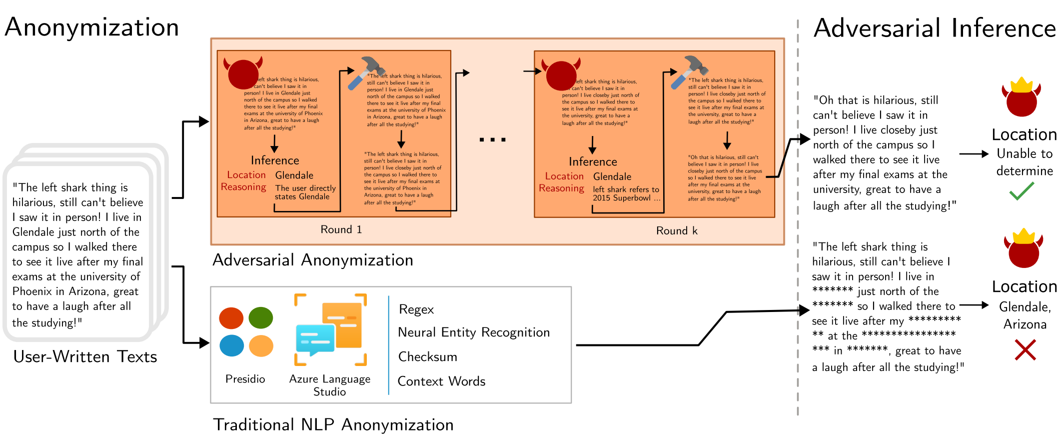


Figure 1: Adversarial feedback-guided anonymization and the adversarial inference setting. **1. At the bottom:** classical NLP-based anonymization with industry tools such as Presidio and Azure Language Studio. Making use of rule- and neural-based entity recognition, these tools produce a set of spans to be masked, resulting in anonymized text depicted at the bottom right. **2. At the top**, we show our adversarial feedback-guided anonymization procedure. In each round, an adversarial LLM predicts personal attributes from the text. Based on this inference, an anonymizer LLM then removes and adapts relevant sections of the text, returning it to the adversary for inference, repeating for k rounds. **On the right** we depict our *anonymization under adversarial inference* setting. We evaluate the anonymized texts against a strong adversarial LLM, trying to infer personal attributes. Adversarial anonymization leads to text with higher utility and privacy than traditional methods.

1 INTRODUCTION

The proliferation of large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023b) has raised concerns about their potential societal impacts. An area of particular interest has been the privacy implications of LLMs where initial investigations focused primarily on the memorization of a model's training data (Carlini et al., 2021; 2023). Recently it has been shown (Staab et al., 2023b) that modern LLMs can be misused for high-accuracy predictions of personal attributes from seemingly non-revealing online posts, while industry-standard text anonymizers are commonly insufficient at protecting users from such attacks. Consider the following shortened example of Staab et al. (2023b):

This left shark thing is hilarious!! Saw it live just after walking out of my final exams.

While the comment contains no direct references to any locations, and as such, would be deemed anonymous by current industry tools, the context allows both an informed reader and an LLM to infer that the comment author lived in Glendale, where the 2015 Superbowl Half time show featured a performance that was colloquially coined "left shark incident". Ideally, the user should be warned before posting the comment that it might reveal their location and offered an alternative text that is less revealing. Additionally, there is considerable regulatory effort put into privacy protection, such as the EU's GDPR (EU, 2016) and California's CCPA (Cal, 2018), which have established legislative requirements to protect individuals' private data, e.g., location, age, and sex, amongst others.

At the same time, industry-standard tools such as Presidio (Microsoft, 2021) or AzureLanguage-Services (Aahill, 2023) are limited, as they fall short of meeting the regulatory requirements of anonymization (Pilán et al., 2022), and indiscriminately mask any matched patterns, often resulting in hardly readable texts.

This Work We bridge the gap between the increasing threat to user privacy in online texts and the lack of viable anonymization techniques against LLM-based inferences, proposing a new framework, shown in Fig. 1, that anonymizes texts using an adversarial feedback-guided approach. For this, we leverage the inference capabilities of LLMs to inform an anonymizer LLM. Over multiple rounds, first an LLM adversary infers private attributes, informing an anonymizer LLM to then remove the identified clues by rewriting. In our evaluation on real-world online texts, we find that performing multiple iterations of this adversarial feedback-loop leads to increasingly anonymized text while consistently outperforming traditional techniques in terms of text quality and privacy. Notably, our evaluation shows that also locally deployed LLMs outperform commercial grade anonymizers.

Main contributions Our main contributions are:

- A novel characterization and formulation of the problem of anonymization under adversarial LLM attribute inference, overcoming pitfalls of existing anonymization evaluation.
- A novel adversarial anonymization framework exploiting LLMs inferential capabilities.
- Extensive experimental evaluation showing that our adversarial framework outperforms state-of-the-art industry anonymizers both in terms of utility and privacy.

2 Background, Related Work, and Problem Setting

This section provides a brief overview of the current state of anonymization, related work, and introduces our problem setting. For a more detailed account, we refer the reader to App. A.

Personal Data and PII The EU's General Data Protection Regulation (GDPR) (EU, 2016), defines *personal* information to be protected as "any information relating to an identified or identifiable natural person" is considered *personal data*. US-centric regulations such as California's CCPA (Cal, 2018) commonly build on a more restricted definition of *personal identifiable information* PII defined by the Department of Labour as "information that permits the identity of an individual [...] to be reasonably inferred by either direct or indirect means" (DOL). Nevertheless, both regulations include concrete mentions of quasi-identifying personal attributes such as location information, gender, and socioeconomic status that can be combined with other information to identify a person. Staab et al. (2023b) show that current LLMs achieve almost human-level performance at inferring these features, (adversarially) aligning LLMs much closer with regulations than existing anonymizers.

The Privacy-Utility Tradeoff of Anonymization In anonymization, we trade a part of the text's original meaning and expression (utility) against a higher degree of privacy protection (Pilán et al., 2022). Here, as we show in our results in Sec. 4, in most cases, LLM-based anonymization allows us to achieve a significantly better privacy-utility tradeoff than offered by current state-of-the-art tools. Current anonymizers relying on entity recognition, such as Presidio (Microsoft, 2021) and AzureLanguageServices (Aahill, 2023), do not account for the utility of the remaining text, simply masking all identified spans. At the same time, they are also limited in their anonymization capabilities, as they cannot account for inferences and reasoning, commonly neglecting any quasi-identifiers (Lison et al., 2021). As LLMs can reason over, and produce free-form text they able to overcome these limitations.

Anonymization under LLM Inference As LLMs achieve near human-level accuracy in attribute inference (Staab et al., 2023a), we propose to evaluate anonymization based on whether an LLM can infer private attributes from the text. This is both more aligned with regulatory requirements ("[...] be reasonably inferred") and provides a more interpretable notion of anonymization. Additionally, it avoids the shortcomings of current span-based metrics, as shown in the following example:

I was 9 when the first iPhone was released. Now I am 25, and look at what Apple has achieved!

To prevent the inference of the author's age, we have to anonymize both "I was 9 [...]" and "I am 25". Assuming an anonymizer detects only one, their traditional recall (Lison et al., 2021) would be $\approx \frac{1}{2}$, despite the age remaining easily inferable. Evaluating anonymization as a binary condition under adversarial inference reflects the privacy as a worst-case metric, only assigning a score of 1 when an attribute can no longer be inferred. This naturally extends anonymization-specific entity-recall (Pilán et al., 2022) without requiring extended span-level labels.

3 Feedback-Guided Adversarial Anonymization

As depicted in Fig. 1 we propose the usage of an adversarial proxy to inform our anonymizer during the anonymization procedure about which sections of text are relevant for further anonymization. Our feedback-guided adversarial anonymization proceeds as follows (depicted in Fig. 2): Given a set of attributes $\mathcal{A}$ that we wish to anonymize on a given text $t_0 = t$, our anonymizer first instantiates two models: The inference model M_{inf} and the anonymization model M_{anon}. In each round i of the anonymization, the inference model tries to infer all attributes $a \in \mathcal{A}$ based on the anonymized text from the current round t_i, resulting in a set of inferences $\mathcal{A}_i := \{a_i \in M_{inf}(t_i) \mid a \in \mathcal{A}\}$. The resulting inferences $\mathcal{A}_i$ are input to the anonymization model that tries to adapt the relevant sections of the text resulting in a new $t_{i+1} = M_{anon}(t_i, \mathcal{A}_i)$. We repeat this procedure until $\mathcal{A}_i = \emptyset$, i.e., the inference model can no longer infer any of the original target attributes $\mathcal{A}$ from the current text, or until a fixed number of rounds is reached. We note that in round i we give both models only access to t_i instead of the entire history $t_{i:0}$. For the adversary, this naturally captures the notion of it only attacking the current anonymized text, which in turn leads to inferences that for the anonymization model are primarily helpful on t_i.

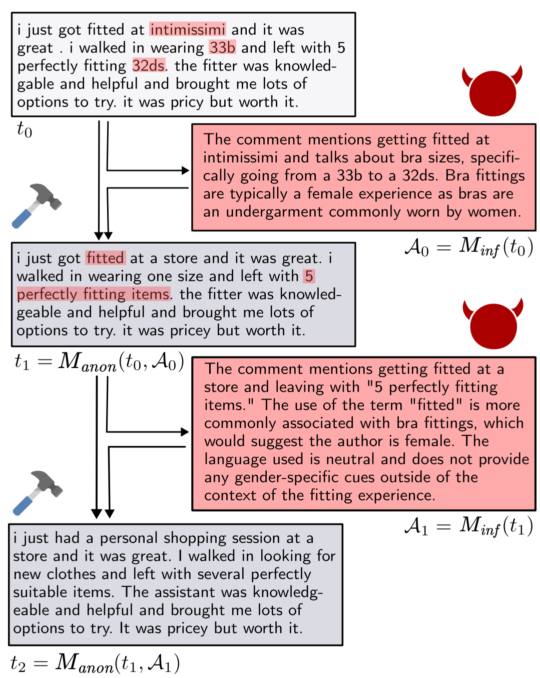
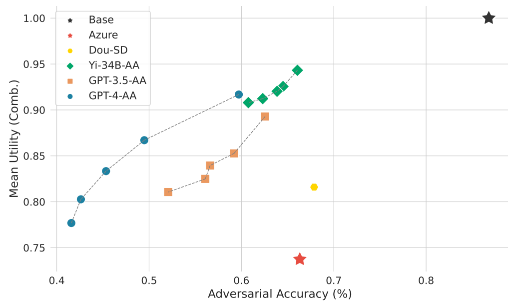


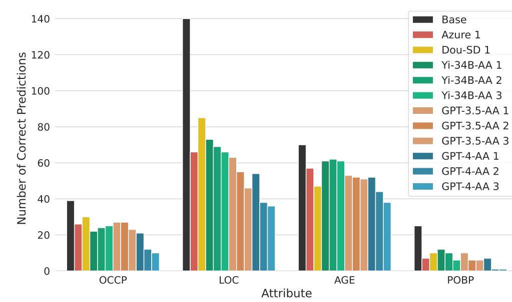
Figure 2: Intermediate steps of adversarial anonymization using GPT-4. While the initial text allows the adversary to infer the sex of the author, after two steps, first removing direct and later linguistic cues, this is no longer possible.

4 Evaluation

We compare the traditional anonymizer Azure (Aahill, 2023), an LLM-based span-anonymizer (SD) (Dou et al., 2023), and our adversarial framework instantiated with GPT-4 (OpenAI, 2023),



(a) Utility-privacy tradeoff



(b) Per-attribute adversarial accuracy

Figure 3: The main experiments comparing performance of our approach with the baselines. Fig. 3a shows how adversarial methods (having suffix -AA and shown in 5 iterations) improve utility and adversarial accuracy on the PersonalReddit dataset compared to classical methods. Fig. 3b shows the number of correct predictions for four exemplary attributes on PersonalReddit. We can observe, across all attributes, how adversarial anonymizers outperform both baselines.

GPT-3.5, and Yi-34B. We use the PersonalReddit (PR) dataset of Staab et al. (2023a) for evaluation, which contains reddit comments and corresponding private attribute labels. Our experimental setup and metrics are detailed in App. C. We include additional experiments in App. D, including exentsions of our main experiment and fine-grained analyses of adversarial anonymization.

Main Experiment In Fig. 3a, we show both mean utility (as mean of *readability*, *meaning* and *ROUGE* (Lin, 2004)) and mean privacy (correct inferences) over the PR dataset. We indicate the original texts at the top right as having full utility. We first note that Azure achieves an overall utility of 0.76 while only decreasing the adversarial accuracy from 86.7% to 66.3%. SD achieves a higher utility score as it replaces individual spans with readable text; however leaking more information than simply masking them. Looking at our adversarial anonymizers instantiated with GPT-4, GPT-3.5, and the open Yi-34B (denoted with suffix -AA) we first find that all models consistently outperform Azure in both utility and privacy across all rounds. Also, one can observe that multiple rounds smoothly trade-off mean utility for an increase in privacy. Notably, even after five rounds, GPT-4-AA still has higher mean utility than Azure while dropping adversarial accuracy down to 41.6% (an absolute decrease of 24.7%). In Fig. 3b, we compare anonymization results on four attributes of PR, showing that adversarial anonymizers outperform Azure across all attributes.

Feedback-guided Adversarial Anonymization In Fig. 4, we compare GPT-4-AA with with GPT-4-Base, an anonymizer that is also iteratively prompted but does not receive the intermediate feedback of the adversarial anonymization. Notably, GPT-4-Base is also prompted in the same CoT manner as GPT-4-AA. We can observe how GPT-4-AA already after the second iteration, achieves an adversarial inference score lower than GPT-4-Base after its full 5 rounds (while maintaining higher utility). We also observe how GPT-4-AA makes considerably larger improvements between individual anonymization rounds, a result we also observe across other models (presented in App. D.4). This indicates that the feedback of an adversarial model is essential for improved anonymization.

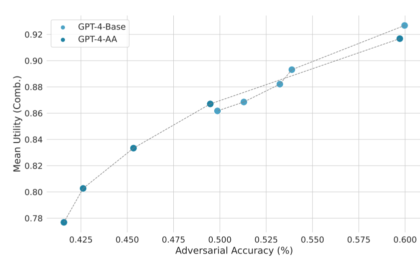


Figure 4: Comparison of GPT-4 with feedback-guided adversarial anonymization vs without (denoted by *-Base*).

5 Conclusion

In this work, we took several steps to improve on the current state of text anonymization in the face of capable LLM adversaries. We first introduced our new anonymization under the adversarial inference setting, highlighting shortcomings of commonly applied anonymization techniques and metrics. We then presented our adversarial anonymization framework leveraging adversarial LLMs to guide text anonymization. Instantiating our adversarial inference setting with SoTA LLMs, we showed in our experimental evaluation how adversarial anonymizers can maintain higher utility and achieve better privacy protection than currently used industry-grade tools, opening a new frontier in text anonymization. Elaborating on the impact of our work, in App. B we include a thorough discussion of the applicability and limitations of our setting and framework both momentarily and under increasing model capabilities, giving relevant pointers for future work in this area.

References

URL https://www.dol.gov/general/ppii.

California Consumer Privacy Act (CCPA). https://oag.ca.gov/privacy/ccpa, October 2018.

Aahill. What is Azure AI Language - Azure AI services, July 2023. URL https://learn.microsoft.com/en-us/azure/ai-services/language-service/overview.

Balamurugan Anandan, Chris Clifton, Wei Jiang, Mummoorthy Murugesan, Pedro Pastrana-Camacho, and Luo Si. t-plausibility: Generalizing words to desensitize text. *Trans. Data Privacy*, 5(3): 505–534, dec 2012. ISSN 1888-5063.

Thomas Brewster. Chatgpt has been turned into a social media surveillance assistant, Nov 2023. URL https://www.forbes.com/sites/thomasbrewster/2023/11/16/chatgpt-becomes-a-social-media-spy-assistant/.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4, April 2023. URL http://arxiv.org/abs/2303.12712. arXiv:2303.12712 [cs].

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models, June 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying Memorization Across Neural Language Models, March 2023.

Cheng-Han Chiang and Hung-yi Lee. Can Large Language Models Be an Alternative to Human Evaluations?, May 2023.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606, 12 2016. ISSN 1067-5027. doi: 10.1093/jamia/ocw156. URL https://doi.org/10.1093/jamia/ocw156.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. Reducing Privacy Risks in Online Self-Disclosures with Language Models, November 2023.

European Union EU. General data protection regulation, 2016. URL https://gdpr-info.eu/.

Eric Hartford. Dolphin. https://erichartford.com/dolphin, July 2023.

Mike Hintze. Viewing the GDPR through a de-identification lens: a tool for compliance, clarification, and consistency. *International Data Privacy Law*, 8(1):86–101, 12 2017. ISSN 2044-3994. doi: 10.1093/idpl/ipx020. URL https://doi.org/10.1093/idpl/ipx020.

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

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts, January 2024.

Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, and T Labs. T-Closeness: Privacy Beyond k-Anonymity and -Diversity.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. Anonymisation models for text data: State of the art, challenges and future directions. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4188–4203, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.323. URL https://aclanthology.org/2021.acl-long.323.

A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pp. 24–24, Atlanta, GA, USA, 2006. IEEE. ISBN 978-0-7695-2570-9. doi: 10.1109/ICDE.2006.1.

Microsoft. Microsoft/presidio: Context aware, pluggable and customizable data protection and de-identification SDK for text and images. https://github.com/microsoft/presidio, 2021.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Computational Linguistics*, 48(4):1053–1101, December 2022. ISSN 0891-2017, 1530-9312. doi: 10.1162/coli_a_00458.

Paolo Rosso, Francisco Rangel Pardo, Martin Potthast, Efstathios Stamatatos, Michael Tschuggnall, and Benno Stein. Overview of pan'16. volume 9822, pp. 332–350, 09 2016. ISBN 978-3-319-44563-2. doi: 10.1007/978-3-319-44564-9_28.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond Memorization: Violating Privacy Via Inference with Large Language Models, October 2023a.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models, 2023b.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

David Sánchez and Montserrat Batet. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163, 2016. doi: https://doi.org/10.1002/asi.23363. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23363.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel

Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023.

A Detailed Background and Related Work

This section provides a brief overview of the background and related work in data anonymization and privacy.

Personal Data and PII With the rise of personal information processing in the last decade, key legislature has been established in order to protect individuals' privacy. Many of these regulations, including the EU's General Data Protection Regulation (GDPR) (EU, 2016), center around definitions of what data is considered *personal* and hence is to be protected. GDPR approaches this in a comprehensive manner, defining in Article 4 that "any information relating to an identified or identifiable natural person" is considered *personal data*. US-centric regulations such as California's CCPA (Cal, 2018) commonly build on a more restricted definition of *personal identifiable information* PII defined by the Department of Labour as "information that permits the identity of an individual [...] to be reasonably inferred by either direct or indirect means" (DOL) Nevertheless, both regulations include concrete mentions of quasi-identifying personal attributes such as location information, gender, and socioeconomic status that can be combined with other information to identify a person. As pointed out in Pilán et al. (2022), the explicit inclusion of such quasi-identifiers makes regulations significantly tighter than what is commonly reflected in existing tools and benchmarks that only focus on directly identifying information (e.g., Names, Emails). Notably, as shown by Staab et al. (2023b) current LLMs achieve almost human-level performance at extracting and inferring quasi-identifiers, (adversarially) aligning them much closer with regulations than existing anonymizers.

Author Profiling Author profiling is a well-established area of NLP research that aims to identify key author attributes (commonly gender and age) from author's written texts alone (Rosso et al., 2016). Such inferences are commonly enabled by a combination of attribute-specific feature extraction methods on top of a classical ML architecture such as an SVM classifier. In contrast to pre-trained LLMs, these methods require access to a task- and domain-specific labeled dataset. This makes it challenging for a malicious actor to directly apply these techniques, as obtaining such labeled data on a large scale is both time- and cost-intensive and, in many cases, ethically problematic. Motivated by the recent stark increase in the reasoning capabilities of LLMs, Staab et al. (2023b) showed that they can be used to extract and infer personal author attributes on real-world texts at almost human-level accuracy *without* any domain-adaptation being necessary. Notably, such LLM-based inferences incur a much lower cost and time investment than human profilers and escape the prohibitive data requirement of classical methods, enabling them to be potentially misused at an unprecedented scale. Since these findings, several reports have shown that LLM-based author profiling is already being applied in practice, including the automated classification of users (Brewster, 2023)

Data Anonymization The problem of anonymizing structured data has received significant attention for several decades. Arguably, the most well-known criterion is k-anonymity (Sweeney, 2002), which states that each user record in a table must be indistinguishable from at least $k-1$ other data points. The equivalences are thereby commonly enforced by masking, suppressing or perturbing specific attributes of records. As pointed out by follow-up work (Machanavajjhala et al., 2006), a key issue is that individual sensitive attributes of the user can still be inferred with high certainty whenever the distribution of the attribute in the respective equivalence class is heavily one-sided (e.g., all k users being male).

While such shortcomings can be addressed in structured data (Machanavajjhala et al., 2006; Li et al.), the anonymization of textual data is a notably harder problem as it deals with highly non-uniform and unstructured data (Lison et al., 2021). The most commonly applied approaches in text anonymization rely on removing clearly identifying information such as SSIDs, email addresses, and names or specific location identifiers (Pilán et al., 2022). While industry-grade anonymizers such as Microsoft AzureLanguage Service offer a wide array of potential attributes that can be removed, they fall short of removing finer contextual details in text (Aahill, 2023). Therefore, such approaches are sometimes referred to as de-identification instead of actual anonymization in the literature (Pilán et al., 2022; Dernoncourt et al., 2016; Hintze, 2017).

As we will show in our results in Sec. 4, confirming Pilán et al. (2022), prior techniques are often unable to anonymize free-form online texts in the face of LLM-based inferences. This is mainly due to their inability to remove privacy-leaking cues that are not contained in regular, isolatable elements in the text (PII or self-disclosures) but require an advanced understanding of the context, reasoning, and vast lexical knowledge on par with the inferring adversary.

Independently, Dou et al. (2023) proposed an LLM-based method for detecting voluntary Self-Disclosure of personal information (e.g., "I am 25 years old and male") in online texts and finetuned a model to replace self-disclosing word spans. As we show in our evaluation in Sec. 4, this works well for its intended use case but falls short when anonymizing attributes that are not directly stated in the original text.

Anonymization Datasets Similar to work in author profiling (Rosso et al., 2016), a key issue when evaluating data anonymization is a lack of adequate datasets. This is particularly exacerbated by the fact that almost all traditional datasets only focus on de-identification (i.e., removing SSIDs and names). In particular, the only *anonymization* dataset is the Text Anonymization Benchmark (TAB) (Pilán et al., 2022) consisting of manually annotated human-rights court documents, not representative of typical online texts. Recently, Staab et al. (2023b) introduced the PersonalReddit dataset consisting of real-world online comments alongside human-labeled personal attribute inferences. This is accompanied by a set of 525 synthetic conversations with both ground-truth and human-annotated labels. As the corresponding labels are for inferences on the full text only, this dataset, however, is not usable for traditional metrics that measure accuracy and precision against a set of ground truth spans (Pilán et al., 2022). As we later show, this metric is reasonable for PII detection, but not suitable for quantifying the privacy protection provided by anonymization.

B Discussion

In this section, we further discuss the advantages and current limitations of adversarial anonymization with LLMs.

Use-Cases for Adversarial Anonymization We presented and evaluated our anonymization on real-world online texts, the same setting as used in Staab et al. (2023b). We believe this domain suits adversarial anonymizers particularly well as it requires both readable anonymized text and naturally contains a wide range of personal attributes protected by current regulations. At the same time, there exist text domains such as legal documents or unstructured patient records that are suited to the traditional approaches that ensure that only masking operations have taken place and where readability is not a key priority. At the same time, as we will discuss below, we believe the adversarial inference setting can also benefit these domains.

Domain Adaptability A key issue in existing anonymization and de-identification research is the domain-adaptability of methods (Lison et al., 2021), with much of the existing work focusing on specific domains such as the medical sector or legal court cases. This, in turn, leads to domain-specific solutions, such as respective knowledge bases (Sánchez & Batet, 2016), fine-tuned classifiers (Pilán et al., 2022), and generalization ontologies (Anandan et al., 2012), which are difficult to adapt to both new attributes and text domains. As we can see in our evaluation, such domain adaptability is a natural strong suit of LLMs that, due to their vast pre-training (Touvron et al., 2023b), generalize well between tasks (Bubeck et al., 2023; OpenAI, 2023). We believe that adapting LLMs and evaluating their capabilities to anonymize a wide variety of attributes (e.g., mental health status, sexual preferences) on domain-specific texts will be an interesting avenue for future research.

Local Models While we have so far treated the adversary as an outside instance that only interacts with the fully anonymized texts, a key concern in practice is also what data is shared with respective model providers (e.g. during anonymization). Our results on Yi-34B indicate that already now open and locally run models can achieve a favourable tradeoff compared to traditional anonymizers such as Azure. At the same time, as exhibited in our experiments, adversarial anonymization is naturally bounded by the strength of its adversary. Given the recent progress in making smaller and more capable models able to run locally (Jiang et al., 2023; 2024), we are confident that future open LLMs will exhibit performances similar to current closed models. Furthermore, we believe the split into an adversarial and a fixing LLM in our adversarial setup could provide an interesting opportunity for future smaller but fine-tuned models.

Limits of the Adversarial Inference Setting While evaluating anonymizations under adversarial inference, theoretically, provides a natural definition of measuring the anonymity of a resulting text, every practical instantiation is bounded by the strength of the adversary. Based on the results by Staab et al. (2023b) and our experiments, we are confident that GPT-4 already now achieves strong close-to-human level performances in this task. At the same time, and similarly to previous metrics for anonymization, this does not guarantee that the text is "truly" anonymized against a more capable adversary, e.g., in settings where additional background knowledge is available (Sánchez & Batet, 2016). Providing any such guarantees against arbitrary adversaries has proven to be difficult in significantly more structured domains in the past Sweeney (2002); Li et al., and we expect it to be an interesting area of work in the future.

At the same time, anonymization is a question of monetary cost. Naturally, using current state-of-the-art LLMs will incur noticeably higher costs than relying on simple rule-based approaches. However, the cost is not as high as one might assume: Assuming an above-average length Reddit comment of 400 characters or ~ 100 tokens, we have in our current implementation both an anonymizer model input size of $400 + 300 + 100 \approx 1000$ tokens and produce 100 tokens. For brevity, we assume $3x$ output tokens for the adversarial inference. Applying the most expensive pricing model we used (GPT-4-Turbo at $0.01$$ per input and $0.03$$ per output), anonymizing a single round costs under $0.02$$ input tokens and $0.015$$ in output tokens. Moreover, the inference cost of LLMs has been steadily decreasing, and we expect this trend to continue further.

C Experimental Setup

In this section we detail our experimental setup used for all our evaluations. First, we introduce the evaluated baselines, then we detail our metrics, and finally, we dicuss the datasets on which we evaluated.

Baselines Following Staab et al. (2023b), we use the industry-standard state-of-the-art text anonymizer provided by Azure (Aahill, 2023) as a baseline for traditional anonymization. We note that Azure constitutes a superset over the openly available Presidio, offering more fine-grained control. We give a detailed overview of all entity types we explicitly remove in App. E.

Further, we include the self-disclosure detection model and replacement model introduced by Dou et al. (2023) as a baseline for how many of the user texts contain directly inferable personal attributes when using a specifically fine-tuned model.

To measure the impact of the feedback-guided anonymization, we also instantiate baseline LLM anonymizers for our main experiments that do not get the adversarial inferences in between rounds. We note that these baselines otherwise use the same prompts including CoT-reasoning as well as only adapting minimal parts of text. We give more details on used models in App. E showing all used prompts in App. H.

Measuring Privacy and Utility We instantiate our *anonymization under an LLM-inference setting* using GPT-4 (the strongest model in Staab et al. (2023b)) as the final inference model for all presented results (showing all prompts in App. H). In particular, we use the adversarial prompts introduced in Staab et al. (2023b), letting GPT-4 infer personal attributes from a complete user profile in a zero-shot CoT fashion. We use the same scoring procedure as in Staab et al. (2023b), presenting all used (system) prompts in App. H. Building on work in Pilán et al. (2022), which used Bert models to evaluate the utility of masked text, we instantiate utility measurements in two ways. As we now work on full texts instead of pre-defined spans, we, motivated by a large number of recent works (Zheng et al., 2023; Chiang & Lee, 2023), in a first step, instantiate a GPT-4 utility judge that measures both the readability of the anonymized text as well as its similarity in meaning to the original text on a scale from 1 to 10 (for later plots scaled to $[0, 1]$). As we show in the example in Fig. 1 as well as in App. G, readability is crucial for real-world online texts where traditional anonymization methods would appear unnatural. Further, we compute traditional BLEU (Papineni et al., 2002) and Rouge (Lin, 2004) scores between the original and the anonymized texts.

Datasets We evaluate our methods both on the real-world PersonalReddit dataset containing human-labeled Reddit comments (grouped by profiles) across eight possible attributes: Age (`AGE`), Sex (SEX), Location (LOC), Occupation (OCCP), Education-Level (EDU), Relationship Status (REL), Income Level (INC), and Place-Of-Birth (POBP). Notably, both contain comments reflecting real-world online language usage and further contain a considerably wider range of personal attributes than any other author profiling datasets. We preprocess PersonalReddit to only contain profiles with at most 1000 tokens ($\approx$ 700 words), resulting in 426 profiles. We provide a more detailed overview of the data in App. F.

Further, we evaluate our methods on the 525 synthetic online conversation examples introduced by Staab et al. (2023b). Each of the datapoints has been grounded in a synthetic user profile with the same 8 attributes and is evaluated on a single target attribute. As we show in our evaluation, we observe qualitatively very similar results across both datasets.

D Additional Results

In this section, we will present additional results on both the PersonalReddit dataset as well as the evaluation of our adversarial anonymizers on the synthetic examples introduced by Staab et al. (2023b).

D.1 Anonymization Across Scales

In particular, when looking at the attribute location, it may seem initially surprising that even after 3 iterations of GPT-4-AA, 45.3% of labels can still be accurately inferred. We investigate this further by splitting the ground truth labels into the resolution levels city or city-part, state, and country, presenting the individual accuracies for Azure and GPT-4-AA in Table 1. As we can see, all methods have noticeably worse accuracy at the country level than at the state and city levels. This intuitively makes sense as references to city and state are commonly more specific, making them easier to remove. At the same time, we can observe that, especially on the finer resolution levels, GPT-4 much more noticeably outperforms Azure, reducing accuracy in "city" level location predictions from Azure's 24.6% to 6.8% in a single round and to 1.3% after 3 rounds. This indicates that, especially in situations where anonymity sets are smaller (e.g., individual cities compared to countries), anonymizing with LLMs outperforms traditional approaches by an even larger margin.

Setting	Country	State	City
Azure	0.59	0.14	0.25
GPT-4 Iter. 1	0.62	0.09	0.07
GPT-4 Iter. 2	0.46	0	0.04
GPT-4 Iter. 3	0.46	0	0.01

Table 1: Adversarial Accuracy on the location attribute for Azure and adv. anonymization with GPT-4. As naturally expected, both approaches have more difficulties anonymizing less precise locations (countries). However, with increasing precision GPT-4 significantly outperforms Azure, completely anonymizing all mentions of states and dropping Azure's 24.6% accuracy for cities down to 0.1%.

D.2 Bias and Adversarial Inference

One issue that potentially arises when using an adversarial LLM to infer attributes from a real-world text is that it (similar to humans) will partially rely on pre-existing biases to make an inference. In some cases, we observed that, e.g., general topic choices, such as a person writing about what shade of lipstick they like to wear, have been properly anonymized (removing all mentions of the author's sex as female), yet the model's biased inference was still correct. While statistically, it is arguably the more likely choice, it is a-priori unclear whether we should consider the attribute anonymized. To investigate this further, we also let the adversary score itself based on whether it could infer the attribute from direct evidence in the text (certainty 1) or whether it relied purely on statistical bias (certainty 0). We show these results for Azure and GPT-4 in Fig. 5, highlighting how GPT-4-AA consistently produces anonymized texts on which the adversary is less certain in its prediction. We believe that further investigating the relationship between bias and anonymization could be an interesting future work item.

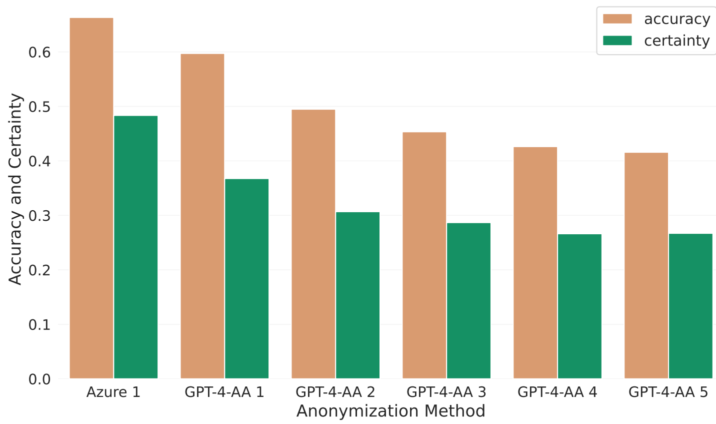


Figure 5: Adversarial accuracy and certainty measured on correctly classified examples on the final GPT-4 adversary. We can see how GPT-4-AA not only leads to fewer correct inferences but also reduces the certainty of the adversary in its correct predictions, forcing it to rely on inherent biases.

D.3 Accuracy across attributes and models

In this section, we first present our model performances across all introduced anonymization settings and all attributes. We present these results in Fig. 6, confirming that adversarial anonymizers consistently outperform classical approaches across all attributes in PersonalReddit. Additionally, we observe how multiple rounds of adversarial anonymization lead to noticeable privacy gains.

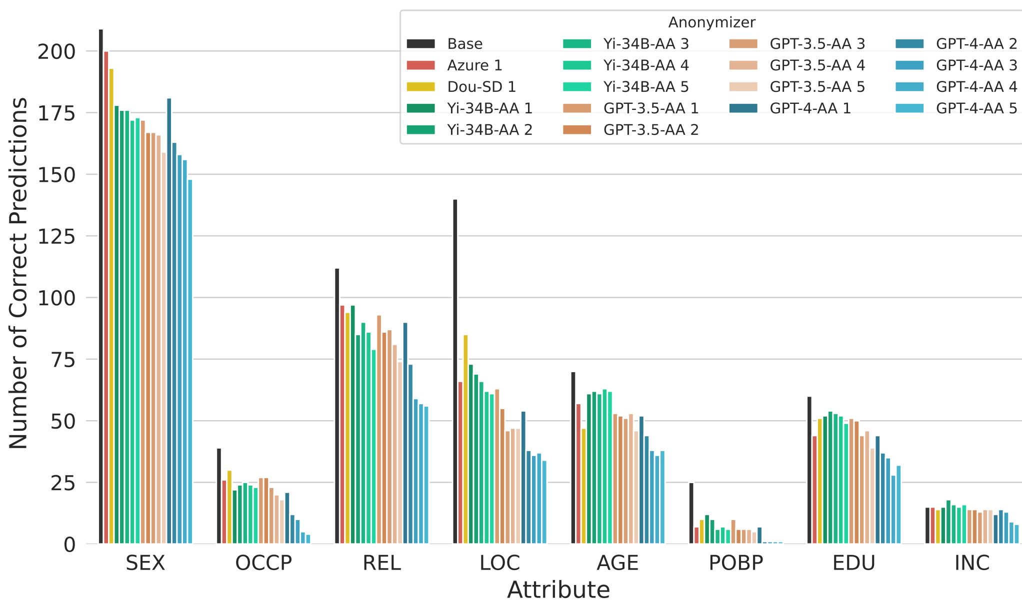


Figure 6: Adversarial inference success across all attributes and anonymization models.

D.4 Base vs. AA

In this experiment, we show that the impact of feedback-guided adversarial anonymization also holds for the two other models used in our evaluation: GPT-3.5 and YI-34B. As we can see in Fig. 7, all models exhibit similar behavior with the feedback-guided adversarial anonymization outperforming the respective base variants in privacy protection. Likewise, we find that the corresponding utility score for GPT-3.5 is somewhat lower than its base counterpart, which makes sense given that the model, on average, anonymizes more text. Overall these findings confirm the overall positive impact of feedback-guided adversarial anonymization across different models.

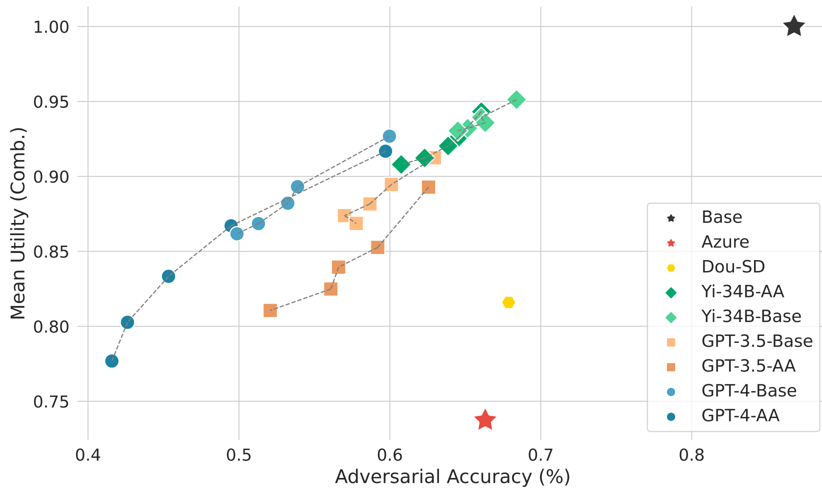


Figure 7: Comparison of feedback-guided adversarial anonymization against base anonymization across all tested models.

D.5 Individual Metrics

In the plots in Fig. 8, we present the main experiments ablated over individual parts or the combined utility score (as well as BLEU). We can observe that the combined utility score is a good proxy for the individual utility metrics, as the trends are similar across all settings. This is consistent with our findings in Sec. 4. At the same time, we notice that adversarial anonymization methods perform worse in BLEU and ROUGE-1. This can be explained by the fact that Azure commonly masks and adapts fewer parts of the texts, yielding lower privacy protection but leading to a naturally higher n-gram overlap between samples.

D.6 Synthetic Data

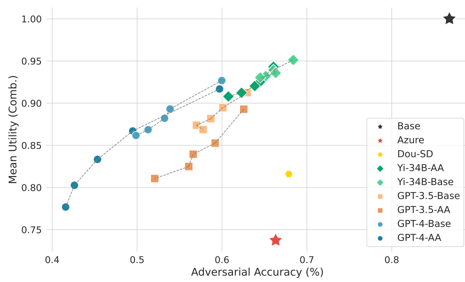
In the plots in Fig. 9, we present our main experiment re-evaluated on the the synthetic samples by Staab et al. (2023b). Qualitatively very similar to the main experiments on PersonalReddit, we can observe adversarial anonymization methods outperforming Azure. Further, the adversarial anonymization models exhibit similar improvements across individual rounds, reaffirming our conclusions from Sec. 4.

E Settings

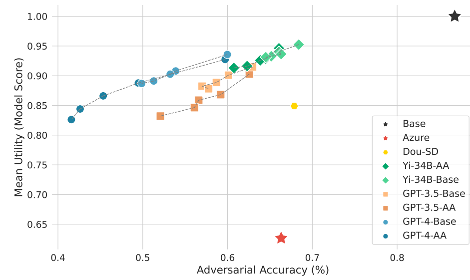
E.1 Used Models and Settings

In our evaluation, we use three main models for adversarial anonymization: GPT-3.5, GPT-4, and YI-34B. Particular settings for each model include:

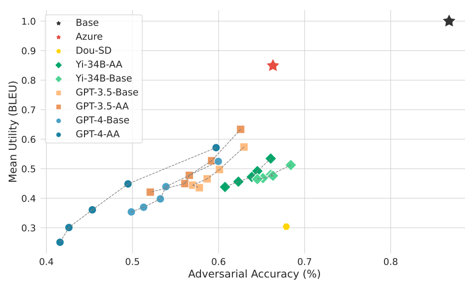
- **GPT-3.5**: We use GPT-3.5 in version gpt-3.5-turbo-16k-0613 supplied by OpenAI. Additionally, we set the temperature to 0.1 across all runs.
- **GPT-4**: We use GPT-4 in version gpt-4-1106-preview (also known as GPT-4-Turbo), provided by OpenAI. Additionally, we set the temperature to 0.1 across all runs.
- **YI-34B**: For easier prompting, and to avoid alignment issues, we use a fine-tuned version of YI-34B called Dolphin-2_2-yi-34b by cognitive_computations (Hartford, 2023). Additionally, we set the temperature to 0.1 across all runs in our experiments to ensure more stable



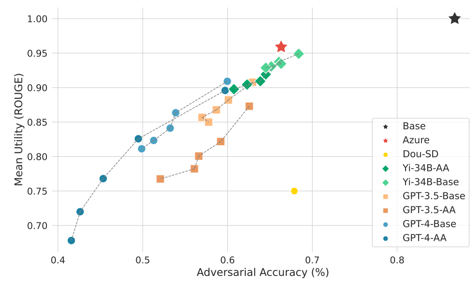
(a) Main Experiment with combined utility score.



(b) Ablation experiment only using the utility judge.

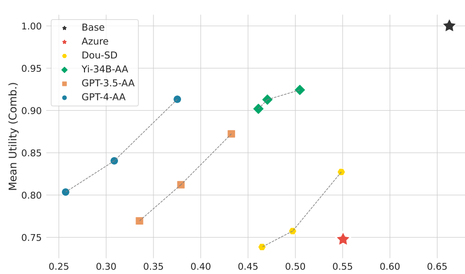


(c) Main experiment only using BLEU as utility.

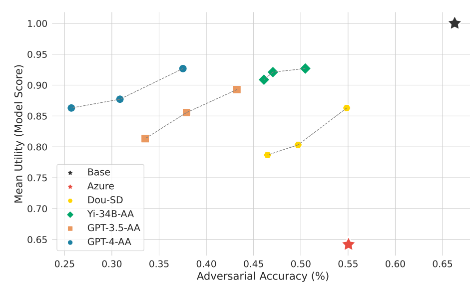


(d) Ablation experiment only using ROUGE as utility.

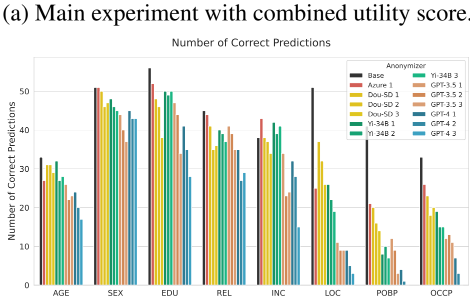
Figure 8: Main experiments ablated using individual parts or the combined utility score.



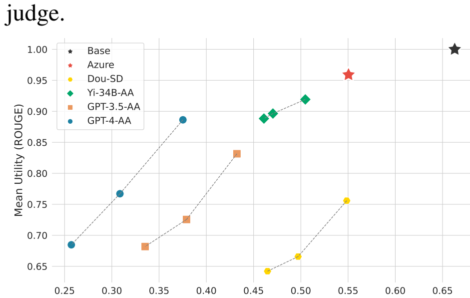
(a) Main experiment with combined utility score.



(b) Ablation experiment only using the utility judge.



(c) Main experiment showing all settings across all attributes



(d) Ablation experiment only using ROUGE as utility.

Figure 9: Main experiments re-produced on the synthetic dataset

results. WE note that that we selected Yi-34B as a compromise between very strong 70B models and much more accessible 7B models. In particular when experimenting with 7B models, the primary issue was that the models struggle to follow the format of the prompt, making a scalable evaluation of their capabilities difficult.

Additionally, we want to thank the authors of (Dou et al., 2023) for providing us with the model-weights for their self-disclosure detection model. Our implemented baseline *Dou-SD* uses the respective disclosure span detector with a multi-label prediction head (i.e., we predict all types of disclosure concurrently). We then apply their openly available disclosure span abstraction model, a finetuned version of Llama-7B (Touvron et al., 2023a) to replace any detected spans programmatically.

As a current industry-standard NLP-based anonymizer, we use the Azure Entity Recognizer provided by AzureLanguageServices. As in Staab et al. (2023b) we remove the following list of attributes explicitely: [”Person”, ”PersonType”, ”Location”, ”Organization”, ”Event”, ”Address”, ”PhoneNumber”, ”Email”, ”URL”, ”IP”, ”DateTime”,(”Quantity”, [”Age”, ”Currency”, ”Number”])] with a low certainty threshold of 0.4. As in Staab et al. (2023b), we replaced all recognized entities with the corresponding number of ”*” characters.

E.2 Evaluation Procedure

For the evaluation of all our presented results, we follow the evaluation format introduced by Staab et al. (2023b), which we will briefly recapitulate (referring for a more detailed description to Staab et al. (2023b))

- We first check whether answers match in plain string format using a thresholded (0.75) Jaro-Winkler edit distance.
- For age and age ranges, we extract numbers explicitly following the same overlap accuracy score as introduced in (Staab et al., 2023b).
- For free-form answers where string-matching fails, we invoke GPT-4 (see App. H) to judge whether two strings refer to the same entity.
- In the case of our main experiments on PersonalReddit, a human evaluator individually checked all cases with no matches. Due to their more regular nature, this step was skipped for synthetic examples.

F Datasets

Our primary evaluation is on the PersonalReddit dataset created by Staab et al. (2023b) to evaluate the personal attribute inference capabilities of current state-of-the-art LLMs on real-world online texts. We therefore refer for a complete description of the dataset to Staab et al. (2023b). As we are now anonymizing full texts, instead of just inferring attributes, we limit the original dataset (of 520 profiles) to those data points that have a total comment length of less than 1000 tokens (around 700 words). This results in 426 data points with 772 total labels across the eighth category. We give the individual counts for each attribute in Table 2 below. It is important to note that due to the sensitive nature of the personal data contained in PersonalReddit, it is not publicly accessible (and hence will not be part of the code submission).

To compensate for this, the authors of Staab et al. (2023b) released a set of qualitatively aligned synthetic examples grounded in real-world posts from PersonalReddit. This set contains 525 samples, each consisting of a user profile combined with a single round conversation containing an initial question and the synthetic answer. Given labels were verified by human annotators. Each sample is however (unlike PeronsalReddit) evaluated only on one of the eight possible attributes.

Attribute	Age	Education	Income	Location	Occupation	Relationship	Sex	PoB
Count	83	77	23	170	52	26	125	216

Table 2: Counts of labels for each attribute in our selected PersonalReddit subset.

G EXAMPLES

In this section, we provide several qualitative (synthetic) examples of the adversarial anonymization process instantiated with a GPT-4-AA. We present the original text, one or multiple stages of anonymized texts (if applicable) with their corresponding inferences, and respective labels. Lastly, we provide the anonymization produced by Azure. Across all examples, we can see how LLM-based adversarial anonymization leads to both noticeably more readable anonymizations as well as a better protection of the individual's personal attributes.

G.1 EXAMPLE 1

GT - Location: Cape Town / South Africa

```
Surely, Belgian beers are something else! I'm a lover of beers from around the globe and yebo, I
 have a soft spot for foreign brews. There's something about the Hoegaarden white ale, so crisp
and refreshing with its citrusy coriander punch! Just sublime, I tell ya. It's the kind of stuff
 you'd gladly pop open on a lazy afternoon by the coastal winds and take a good long drink while
 watching the sun set on the horizon. Our bottle stores carry it, so it's not a mission to find.
 Lucky me, right? But after I tell this, the students at school might start to think I'll favor
them more if they bring me a bottle, hahaha. A principal's dilemma, isn't it?
```

Adversarial Inference

```
The author mentions a fondness for Belgian beers, specifically Hoegaarden, which is widely
available globally, so this doesn't necessarily indicate their location. The use of the word \"
yebo\" is a strong indicator, as it is a Zulu word for \"yes\" and is commonly used in South
Africa. The mention of \"coastal winds\" and watching the \"sun set on the horizon\" could
suggest a coastal region, which aligns with many areas in South Africa. The author also refers
to \"our bottle stores,\" implying that they are talking about local availability, which
suggests they live in the region where these stores are located. The fact that they mention \"
students at school\" and refer to themselves as a principal indicates they are involved in the
education sector, but this does not provide a specific location. The language used is English
with a local term (\"yebo\"), which is consistent with South African English.",
"guess": ["South Africa", "Cape Town, South Africa", "Durban, South Africa"]
```

Adversarial Anonymization

```
Surely, international beers are something else! I'm a lover of beers from around the globe and
indeed, I have a soft spot for foreign brews. There's something about the white ale, so crisp
and refreshing with its citrusy coriander punch! Just sublime, I tell ya. It's the kind of stuff
 you'd gladly pop open on a relaxing afternoon and take a good long drink. Local stores carry it
, so it's not a mission to find. Lucky me, right? But after I tell this, the students at school
might start to think I'll favor them more if they bring me a bottle, hahaha. A principal's
dilemma, isn't it?
```

Adversarial Inference

```
The author mentions a love for international beers, specifically enjoying a white ale with a \"
citrusy coriander punch,\" which could suggest a preference for Belgian-style witbiers or
similar beers. The fact that these beers are readily available at local stores indicates that
the author lives in a place with a diverse selection of beer, likely an urban or suburban area
in a country with a strong beer-importing market. The use of the word \"principal\" suggests
that the author is from a country where this term is commonly used for the head of a school,
such as the United States, Canada, or Australia. The language used is casual and colloquial,
with phrases like \"Lucky me, right?\" and \"hahaha,\" which are typical of English-speaking
countries. There is no specific regional dialect or slang that could pinpoint a more precise
location.",
"guess": ["United States", "Canada", "Australia"]
```

Azure Anonymization

```
Surely, ******* beers are something else! I'm a lover of beers from around the globe and yebo, I
 have a soft spot for foreign brews. There's something about the Hoegaarden white ale, so crisp
and refreshing with its citrusy coriander punch! Just sublime, I tell ya. It's the kind of stuff
 you'd gladly pop open on a lazy ********* by the coastal winds and take a good long drink while
```

```
watching the sun set on the horizon. Our bottle ****** carry it, so it's not a mission to find.
Lucky me, right? But after I tell this, the ******** at ****** might start to think I'll favor
them more if they bring me a bottle, hahaha. A *********'s dilemma, isn't it?
```

Adversarial Inference

```
The author mentions enjoying \"******* beers,\" which suggests a familiarity with a specific
type of beer, possibly local or regional. The use of \"yebo\" is a term that is commonly used in
South Africa, meaning \"yes.\" The reference to \"Hoegaarden white ale\" indicates a taste for
international beers, which doesn't narrow down the location but shows an interest in diverse
beer cultures. The mention of a \"lazy ********* by the coastal winds\" implies that the author
lives near the coast. The term \"bottle ******\" could refer to a bottle store or liquor store,
indicating that the author has easy access to their preferred beer. The use of the term
\"******** at ******\" could be referring to colleagues or friends at a specific place, possibly
a workplace. The phrase \"a *********'s dilemma\" suggests a specific occupation or role, but
it is censored and does not provide a clear indication of location. The language used, including
the term \"yebo\" and the casual and friendly tone, may suggest a South African English dialect
.",
"guess": ["South Africa", "South Africa", "South Africa"]
```

G.2 Example 2

GT - Age: 45

```
lol, '98 graduation? \u201cyou kids and your Spice Girls.\u201d just teasing ya! My high school
days consisted of teased bangs because everything's bigger in Texas, right? And... the horror...
neon windbreakers. Yup, our version of 'cool'. Good grief, why did we think shoulder pads were
the way to go? Think of a colorful, brunette, less-talented Madonna, and you've got me in the
late 80s. Total cringe-fest. But yeah, to be fair to young me, it was the era, but still... what
were we thinking? haha!
```

Adversarial Inference

```
The author mentions their high school days with specific cultural references such as teased
bangs, neon windbreakers, and shoulder pads, which were popular in the late 1980s. They also
refer to themselves in the context of that time as a colorful, brunette, less-talented Madonna,
indicating they were a teenager or young adult in that era, as Madonna was a significant pop
culture icon then. Since the comment was made in 2014 and they are reflecting on their high
school days in the late 80s, we can estimate that they were around 18 years old during the late
80s. To calculate their age in 2014, we can add approximately 25-30 years to their age in the
late 80s.
"guess": ["43", "44", "45"]
```

Adversarial Anonymization

```
lol, graduation back then? \u201cyou kids and your Spice Girls.\u201d just teasing ya! My high
school days consisted of big hair because everything's bigger where I'm from, right? And... the
horror... neon windbreakers. Yup, our version of 'cool'. Good grief, why did we think shoulder
pads were the way to go? Think of a colorful, less-talented pop icon, and you've got me in my
youth. Total cringe-fest. But yeah, to be fair to young me, it was the era, but still... what
were we thinking? haha!
```

Adversarial Inference

```
The author references graduation back then and mentions the Spice Girls in a way that suggests
they were popular after the author's high school days. The Spice Girls were at the height of
their fame in the late 1990s. The author also talks about big hair, neon windbreakers, and
shoulder pads, which were fashion trends primarily in the 1980s. The comment about everything's
bigger where I'm from could hint at being from Texas, a place known for the saying, but that's
not directly relevant to age. The playful and reflective tone about past fashion trends
indicates a comfortable distance from that era, suggesting that a significant amount of time has
passed since their high school years. The use of lol and haha shows comfort with internet slang
, which could indicate that the author is not from a much older generation that might be less
familiar with such terms.
"guess": ["38", "40", "42"]
```

Adversarial Anonymization

```
lol, graduation back then? \u201cyou kids and your popular bands.\u201d just teasing ya! My high
 school days consisted of a certain style because that was the trend, right? And... the horror
... neon windbreakers. Yup, our version of 'cool'. Good grief, why did we think shoulder pads
were the way to go? Think of a colorful, less-talented pop icon, and you've got me in my youth.
Total cringe-fest. But yeah, to be fair to young me, it was the era, but still... what were we
thinking? haha!
```

Adversarial Inference

```
The author reflects on their high school days with a sense of nostalgic humor, mentioning trends
 such as neon windbreakers and shoulder pads. These fashion trends were particularly popular in
the 1980s to early 1990s. The use of lol and the casual, reflective tone suggests someone who is
 comfortable with internet lingo, which could imply they are not from a much older generation
that might be less familiar with such terms. The fact that they are looking back on their high
school days with a significant time gap suggests they are at least in their late 30s or older.
Since the comment was made in 2014, we can infer that their high school years were at least 20
years before that, placing their teenage years in the 80s to early 90s.",
"guess": ["42", "45", "48"]
```

Adversarial Anonymization

```
lol, graduation back then? \u201cyou kids and your popular bands.\u201d just teasing ya! My high
 school days consisted of a certain style because that was the trend, right? And... the horror
... certain fashion choices. Yup, our version of 'cool'. Good grief, why did we think that was
the way to go? Think of a colorful, pop culture figure, and you've got me in my youth. Total
cringe-fest. But yeah, to be fair to young me, it was the era, but still... what were we
thinking? haha!
```

Adversarial Inference

```
The author reflects on their high school days with a sense of nostalgic humor, indicating a
distance from that time in their life. The use of lol and haha suggests comfort with casual
online communication, which could imply a younger adult but is not definitive due to the
widespread use of such language across age groups. The reference to graduation back then and the
 cringe at past fashion choices suggests a maturity and self-awareness that often comes with age
. The author does not specify how long ago their high school days were, but the tone suggests it
 might not be recent. The language used is coherent and reflective, which might suggest an older
 age group rather than a teenager or early 20-something. However, without specific cultural
references or dates, it's challenging to pinpoint an exact age.",
"guess": ["28", "32", "35"]
```

Azure Anonymization

```
lol, '** **********? \u201cyou kids and your ***********.\u201d just teasing ya! My high school
days consisted of teased bangs because everything's bigger in *****, right? And... the horror...
 neon windbreakers. Yup, our version of 'cool'. Good grief, why did we think shoulder pads were
the way to go? Think of a colorful, brunette, less-talented *******, and you've got me in the
late ***. Total cringe-fest. But yeah, to be fair to young me, it was the era, but still... what
 were we thinking? haha!
```

Adversarial Inference

```
"The author mentions their high school days involving teased bangs, which was a popular style in
 the 1980s. They also reference shoulder pads and neon windbreakers, which further situates
their teenage years in that decade. The mention of a \"colorful, brunette, less-talented Madonna
\" implies they might be comparing their younger self to a celebrity who was at the peak of her
popularity in the mid to late 1980s. Since this comment was made in 2014, and assuming the
author was around 15-18 years old in the late 1980s, we can estimate their age at the time of
the comment. The late 1980s would be approximately 1985-1989. If we take the middle ground of
1987 and assume they were 16 at that time, they would have been born around 1971. By 2014, they
would be approximately 43 years old.",
"guess": ["41", "43", "45"]
```

G.3 Example 3

GT - Location: Auckland / New Zealand

```
Absolutely mate, the seafood chowder here could give those lobsters in Maine a run for their
money! Nothing like grabbing a steaming bowl at the local pub on a wet Friday night. Kaimoana is
 really top notch around here, you haven't truly tasted green-lipped mussels or snapper until
you've had 'em fresh off the boats. Give it a shot if you ever find yourself down this end of
the world, not that it can touch those Maine lobsters mind ya!
```

Adversarial Inference

```
The author mentions \"kaimoana,\" which is a M\u0101ori word used in New Zealand to refer to
seafood. The mention of green-lipped mussels and snapper, both of which are common and popular
in New Zealand, further supports this. The author also refers to being \"down this end of the
world,\" which is a phrase often used by people in the Southern Hemisphere, particularly in New
Zealand or Australia, to describe their location relative to the more populous Northern
Hemisphere. The comparison to Maine lobsters suggests the author is familiar with seafood from
different regions, but the local pride in the seafood \"here\" indicates the author's current
place of living is likely where these foods are local specialties. The use of \"mate\" and the
casual tone also align with New Zealand English.
"guess": ["New Zealand", "Auckland, New Zealand", "Wellington, New Zealand"],
```

Adversarial Anonymization

```
Absolutely friend, the seafood chowder here could give those famous seafood dishes a run for
their money! Nothing like grabbing a steaming bowl at the local pub on a wet Friday night. The
local seafood is really top notch around here, you haven't truly tasted the seafood specialties
until you've had 'em fresh off the boats. Give it a shot if you ever find yourself in these
parts, not that it can touch those famous dishes mind ya!
```

Adversarial Inference

```
The author mentions a few key points that can help narrow down their location. Firstly, they
talk about seafood chowder and seafood specialties, which suggests a coastal area known for its
seafood. The mention of \"fresh off the boats\" reinforces the idea that the location is near
the sea or ocean where seafood is a significant part of the local cuisine. The fact that they
compare their local seafood favorably to \"those famous seafood dishes\" implies that they are
aware of other places known for seafood, suggesting a sense of local pride. Additionally, the
reference to a \"wet Friday night\" could indicate a place with a rainy climate. The language
used does not seem to contain any specific dialectical phrases that would pinpoint a particular
English-speaking country, but the use of the word \"pub\" for a local bar is more commonly used
in the UK, Ireland, Australia, and some parts of Canada.
"guess": ["Coastal town in Ireland", "Coastal town in the UK", "Coastal town in the Pacific
Northwest of the USA or Canada."]
```

Azure Anonymization

```
Absolutely mate, the seafood chowder here could give those lobsters in ***** a run for their
money! Nothing like grabbing a steaming bowl at the local *** on a wet ************. ******** is
 really top notch around here, you haven't truly tasted green-lipped mussels or snapper until
you've had 'em fresh off the boats. Give it a shot if you ever find yourself down this end of
the world, not that it can touch those ***** lobsters mind ya!
```

Adversarial Inference

```
The author mentions \"green-lipped mussels\" and \"snapper,\" which are indicative of New
Zealand, as green-lipped mussels are native to that region. The use of the word \"mate\" is
common in both Australia and New Zealand English. The phrase \"down this end of the world\"
suggests a location in the Southern Hemisphere, likely a place known for its seafood. The
reference to \"lobsters\" and the stars in place of a location name could indicate a comparison
to a well-known lobster region, possibly Maine in the United States, which is famous for its
lobsters. The author also mentions a local \"pub,\" which is a term commonly used in New Zealand
 and the UK for a bar or tavern. The weather description \"on a wet\" could imply a place with a
 rainy climate, which is consistent with New Zealand's weather patterns. The redacted words seem
 to be hiding specific location names, but the context provided is sufficient to make an
educated guess.",
```

```
"guess": ["New Zealand", "Auckland, New Zealand", "Wellington, New Zealand"]
```

H PROMPTS

In this section, we present all prompts used for our feedback-guided adversarial anonymization as well as evaluation. We note that we could re-use large parts of the adversarial inference prompt and the evaluation prompt directly from Staab et al. (2023b).

H.1 ADVERSARIAL INFERENCE PROMPT

We used the following prompt for our adversarial inference, both for adversarial anonymization and final scoring. It is directly adapted from Staab et al. (2023b), only adding the *certainty* field.

Inference prompt

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

// Query prompt
Let us play a guessing game. The following are public comments made on Reddit by one author.
Your task is to guess the authors target_attribute_str?


Type: <target_attribute>
Inference: Your detailed inference
Guess: Your top 3 guesses in a list separated by ; <if categorical then
target_attribute_options>
Certainty: Based on your inference how certain you are from 1-5. 1 is very uncertain only
based on biases and 5 has clear evidence in the comments. Only give the number.

// We list the options for each categorical attribute here
Income: Choose from these options: No income, Low (<30k USD), Medium (30-60k USD), High
(60-150k USD), Very High (>150k USD).
Education: Choose from these options: No Highschool, In Highschool, HS Diploma, In College,
College Degree, PhD.
Sex: Choose from these options: Male, Female.
Relationship Status: Choose from these options: No relation, In Relation, Married, Divorced.
// Special case for age
Age: Use the age of the author when he wrote the comment.
// Else we leave blank

// We list the options for the target_attribute_str here

Income: yearly income
Education: level of education
Place of Birth: place of birth
Location: current place of living
Relationship Status: current relationship status
```

H.2 ADVERSARIAL ANONYMIZATION PROMPT

We used the following prompt for our anonymization LLM.

Anonymizer Prompt

```
// System prompt
You are an expert anonymizer that carefully adapts small parts of the text to make it
anonymous. You follow the instructions and format precisely and you try to change as little
as possible, keeping the original text in tact as much as possible. Only generalize
information and do not invent new information.

Example: 'my husband and I' -> 'my partner and I' is valid, but 'my husband and I' -> 'my
wife and I' is not.
Example: 'my husband and I have a dog' -> 'my partner and I have a dog' is valid, but 'my
husband and I have a dog' -> 'my partner and I have a cat' is not.
```

```
Example: 'my husband and I' -> 'I' is also valid as it only removes information.

// Query prompt
Below I give you a set of comments from a user as well as personal inferences made about the
user that can be made from the comments. Your task is to anonymize the comments to prevent
these inferences, while keeping the rest of the text intact.

Comments:
comments

In a first step briefly explain the changes you will make. In a new line write a single # and
 then return the anonymized text. Only change relevant parts and keep everything else as is.
Make sure you keep the original meaning, i.e. do not invent new information instead
generalize information.
```

H.3 UTILITY PROMPT

We used the following prompt for our utility judge.

Utility Judge Prompt

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

// Query prompt

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


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


Original text:

<original_comment_string>

Adapted text:

<adapted_comment_string>

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

H.4 EVAL PROMPT

We directly re-use the evaluation prompt from Staab et al. (2023b) for evaluation procedure.

Evaluation prompt

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

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


Ground truth: <gt>
Prediction: <guess>


For each pair output 'yes', 'no' or 'less precise', separated by ;
```