Text Re-identification Evaluation with Retrieval-Augmented Generation

Benet Manzanares-Salor[1][0000−0001−7254−2560] and David Sánchez[1][0000−0001−7275−7887]

Universitat Rovira i Virgili,
Department of Computer Engineering and Mathematics,
CYBERCAT-Center for Cybersecurity Research of Catalonia,
ComSCIAM-Center for Computational Science and Applied Mathematics,
Av. Paisos Catalans 26, Tarragona, 43007, Spain
{benet.manzanares,david.sanchez}@urv.cat

Abstract. Textual data comprise a large part of web content and serve as the foundation for many online services. However, when such data contain personally identifiable information, their disclosure can create significant privacy risks. A common mitigation is anonymization, which seeks to unlink an individual's identity from their associated information. Nevertheless, because most text anonymization techniques lack formal privacy guarantees, the residual re-identification risk must be empirically evaluated *ex post*. Evaluations efforts in the literature typically rely on comparisons with human-produced anonymizations—an approach that is costly, subjective, error prone, and fails to capture the nuanced strategies a determined adversary might employ to re-identify anonymized texts. To address these limitations, we propose an automatic text re-identification evaluation framework grounded in retrieval-augmented generation (RAG) and large language models (LLMs) that simulates a powerful but realistic adversary attempting to link anonymized texts with the identities they refer to. We demonstrate its effectiveness by evaluating the broadest diversity of text anonymization methods examined in the privacy literature to date. Empirical results reveal that most methods pose substantially greater privacy risks than those estimated through human-based evaluations, indicating that such evaluations are becoming increasingly unreliable in the era of generative AI.

Keywords: Privacy · Text anonymization · Re-identification · Large Language Models · Retrieval Augmented Generation.

1 Introduction

The public disclosure of sensitive personal data conflicts with privacy regulations, including the GDPR[6]. Anonymization techniques aim to conceal identifying information, ensuring that data can no longer be unequivocally linked to the individuals they concern. Because anonymized data are not regarded as personal and thus fall outside the scope of privacy legislation such as the GDPR, they can be freely shared or released online.

1.1 Related work

Anonymization of structured (tabular) databases is a mature research area supported by a broad range of methods that provide *ex ante* privacy guarantees [3], including k-anonymity [22] and ϵ-differential privacy [5].

In contrast, practical anonymization methods for *textual* data remain mainly heuristic and therefore lack formal guarantees [11, 18]. For such methods, the achieved protection against re-identification must be empirically evaluated *ex post*. The standard evaluation approach involves comparing anonymized texts to *human annotations*—which serve as a ground truth—and assessing privacy protection using the *recall metric*—which quantifies the proportion of re-identifying terms correctly masked during anonymization [17, 21, 15].

However, this evaluation methodology has several limitations. First, it relies on human judgment, which is subjective, error-prone, and time-consuming. Second, it assumes the existence of a single ground truth, which is unrealistic in text anonymization, where multiple combinations of masked terms can effectively prevent re-identification. Finally, the recall metric treats all text spans as equally important, overlooking their varying impact on identity disclosure.

These evaluation limitations cast doubt on the results produced by text anonymization methods, especially given the wide range of information sources available online that could facilitate re-identification.

1.2 Contributions

We present two main contributions. First, we propose an automatic and objective privacy metric for text anonymization evaluation. By leveraging a large language model (LLM) in combination with retrieval-augmented generation (RAG) [9], our metric operationalizes realistic adversarial capabilities and accounts for the dynamic, continuously expanding nature of Web-scale data, which is the usual source of background knowledge for re-identification.

Second, using our metric, we perform a comprehensive evaluation that encompasses the broadest variety of text anonymization methods examined in the literature to date. Our analysis considers not only classical approaches grounded in *named entity recognition* (NER) and *privacy-preserving data publishing* (PPDP) principles, but also recent LLM-based techniques. By comparing our metric's assessment with the standard evaluation based on the recall metric against manual anonymization, we show that the latter consistently and substantially underestimates re-identification risks across all methods. These findings highlight the limitations of human-based evaluations, which are increasingly unreliable against realistic, AI-powered adversaries.

The remainder of the paper is structured as follows. Section 2, introduces and formalize our risk assessment approach. Section 3 reports and discusses the empirical findings derived from applying our metric to various anonymization techniques, contrasting these results with the standard recall metric. Finally, Section 4 gathers the conclusions and outlines potential directions for future research.

2 Re-identification assessment with RAG

Our metric draws inspiration from established privacy evaluation methods designed for structured databases. In particular, *record linkage* [1] quantifies the re-identification risk of an anonymized database by attempting to link anonymized records with a background knowledge (BK) database containing public *identified* records about the individuals represented in the anonymized database. Because the exact set of protected individuals is typically unknown to the adversary, the BK database is ideally a superset of potential individuals, ensuring coverage of all plausible candidates. In practice, however, this superset may be incomplete, thereby imposing an upper bound on the achievable re-identification risk.

The linkage process relies on the similarity between the attributes shared by both the public and anonymized databases. A successful link connects an anonymized record (including potentially confidential attributes) to a known identity, thereby resulting in effective re-identification. The proportion of successful links quantifies the *re-identification risk* of the anonymized database.

Applying this methodology to textual documents poses substantial challenges. Unlike structured attributes, linguistic expressions in plain text documents are context-dependent, making attribute-level comparisons difficult. Consequently, any word, phrase, or combination of terms may serve as an identifying feature [7, 13, 14].

Our approach can be viewed as an adaptation of record linkage for structured databases to textual corpora, where (combinations of) text spans act as attribute values, documents represent records, and the proposed RAG-based framework links anonymized documents to identified publicly available BK.

More specifically, RAG [9] offers a principled framework for integrating external knowledge into pretrained LLMs. It retrieves task-relevant passages from a curated corpus and incorporates them into the model's prompt, enabling knowledge-grounded responses without (costly) fine-tuning. The RAG framework consists of three key components: (i) a *vector database*, storing domain-specific information; (ii) a *retriever*, efficiently selecting the most relevant text for a query; and (iii) a *reader*, an LLM that generates the final output based on the retrieved content.

For our risk assessment, the retriever selects the texts from the BK vector database (along with their associated identities) that are semantically closest to each anonymized document. The reader then evaluates these candidates to determine which one best matches the anonymized document, thereby linking the corresponding BK identity to it. Finally, the linked identities are compared with the true identities of the anonymized documents to compute the overall re-identification risk of the corpus. Figure 1 illustrates this pipeline, which is detailed in the following subsections.

The re-identification process leverages the fact that anonymized texts may still preserve information associated with the original individual. On the other hand, BK documents, which contain publicly available information about known individuals, may contain overlapping information. A protected document is then linked to a BK identity based on the similarity between such preserved informa-

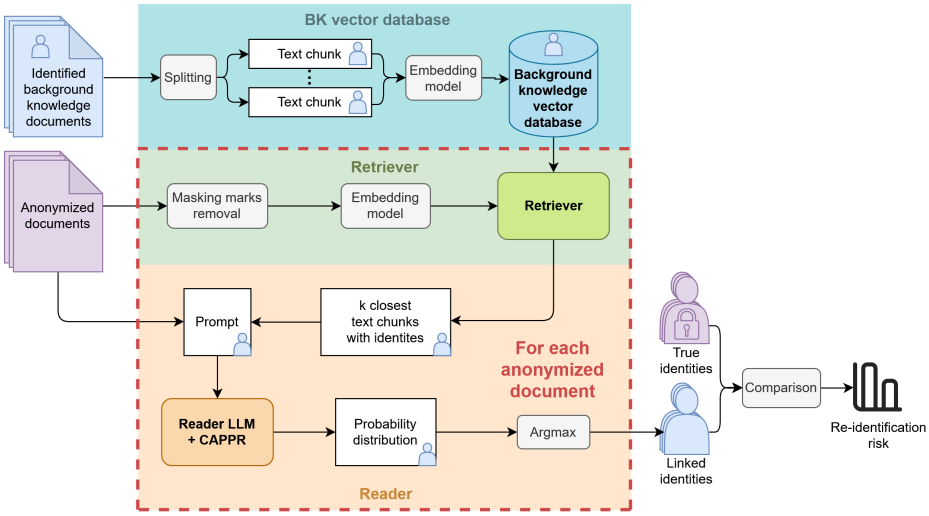


Fig. 1. Proposed re-identification evaluation pipeline

tion and that present in the (identified) BK corpus. When this linked identity matches the originally protected individual, it constitutes a re-identification.

2.1 Background knowledge vector database

The collection of identified public BK documents forms the curated, in-domain corpus for the RAG framework. Each document, associated with a known identity, is represented in plain text and is expected to contain information that overlaps with the anonymized documents. Identities are propagated throughout the re-identification pipeline, as indicated by the blue profile icons in Figure 1.

For example, given a collection of anonymized medical reports from hospital patients, the public social media posts authored by residents of the same city could provide a suitable source of BK. The information used for linkage in this scenario can include demographic data, habits, visited locations, and reported incidents.

For the subsequent retrieval, the BK is stored in a *dense* vector database using the following procedure. First, each document is split into text chunks with a predefined maximum length, allowing the retriever to focus on spans containing relevant information and reducing off-topic content for the reader.

Next, each text chunk is encoded using an embedding model to obtain a dense vector. This representation captures the semantic content of the chunk, that is expected to align more closely with the anonymized documents (which may originate from a different data distribution) than sparse vectors based on term frequencies. Finally, the embeddings, along with their text chunks and associated identities, are stored in the vector database.

The BK construction process (top left in Figure 1) is performed once for the entire re-identification risk assessment.

2.2 Retriever

Given an anonymized document, we first preprocess it by removing any tags introduced during anonymization to mask sensitive content (*e.g.*, "***", "SENSITIVE", named entity types such as "PERSON", "LOCATION", etc.). These masking marks do not convey semantic content and can in fact hinder similarity search, which should focus on the unmasked (*i.e.*, potentially missed) information that overlaps with the BK. Following preprocessing, the document is encoded using the same embedding model employed for the BK vector database, and the resulting vector is supplied to the retriever.

To perform the retrieval, we employ a similarity search system, which enables efficient nearest-neighbor lookups over dense vector representations. The retriever returns the k text chunks from the BK vector database that are the most semantically similar to the anonymized document, along with the BK identity (*e.g.*, an individual's name) associated with each chunk. These identities constitute the pool of candidates for re-identification. In some cases, fewer than k unique individuals might be retrieved if multiple chunks originate from the same individual. If all k chunks map to a single individual, the subsequent reader stage is omitted, as that individual can be selected as the linked identity.

2.3 Reader

We next use a reader LLM to determine which of the retrieved candidate identities best corresponds to the anonymized document.

The LLM is provided with a dedicated prompt that contains the retrieved text chunks (grouped by identity), the list of candidate identities, and the anonymized document itself. In contrast to the retrieval phase, the entire anonymized document (including its masking marks) is supplied to the LLM. The rationale is that, at this stage, the LLM acts as a discriminative judge rather than relying solely on semantic similarity, and the masking marks signal which data have been protected, which are cues that can be leveraged for re-identification. For example, if the protected document repeatedly references a masked "ORGANIZATION," this recurring pattern may strongly suggest a particular candidate identity.

For prediction, we adopt *completion after prompt probability* (CAPPr)[1]. This technique computes the probability distribution of the LLM over a predefined set of possible completions, which, in our case, are the candidate identities. Unlike deterministic greedy generation (*e.g.*, with $temperature = 0$ or $top\text{-}k = 1$), CAPPr removes the need for post-processing the generated output, such as resolving formatting ambiguities, by directly producing normalized probabilities over all candidates based on the model's token-level predictions.

[1] `https://github.com/kddubey/cappr`

The candidate with the highest probability (obtained with an `argmax` function) is selected as the linked identity for the given anonymized document. Once this process is executed across all anonymized documents in the corpus to evaluate, we can compare the linked identities with the true ones to quantify the overall re-identification risk.

2.4 Scalability and flexibility

As with any solution that relies on LLMs, computational cost may be a concern. This could be at odds with GDPR's Recital 26[2], which emphasizes that re-identification risk assessments should be based on a realistic (rather than idealized or unlimited) estimate of the resources an attacker could reasonably devote.

In this respect, the design of our framework is efficient and inherently scalable. First, RAG significantly improves the accuracy of the results with minimal cost, enabling success with reduced-size/efficient LLMs, which makes the attack more attractive. Second, the construction of the BK vector database, retrieval, and reader predictions is performed independently for each document, achieving scalability through parallel processing at the document level.

Moreover, the system supports efficient incremental updates to the BK, which are frequent in dynamic environments such as the Web. When new BK documents are added, the vector database can be updated with minimal computational cost. Retrieval is then re-executed for all anonymized document embeddings, which can be cached to avoid redundant computations. Finally, the reader LLM predictions, which are the most computationally intensive step, are recalculated only for documents where the k retrieved text chunks have changed. It is important to note that, when the proportion of new BK documents is small relative to the existing corpus, most predictions remain valid, minimizing the need for recomputation.

Overall, as shown in the next section, our framework can be executed on mid-tier hardware with very reasonable runtimes.

3 Experiments

We tested our metric by evaluating a wide variety of text anonymization approaches [20, 11] on a domain-independent corpus within a realistic threat scenario. Code and data are available at `https://github.com/BenetManzanaresSalor/Re-IdentificationRiskRAG`.

Experiments were performed on a consumer-grade system equipped with an NVIDIA RTX 4080 GPU (16GB VRAM), an Intel i7-12700 CPU, and 32GB of DDR4 RAM. The software environment included PyTorch 2.9.0, CUDA 12.6, and HuggingFace's Transformers 4.57.1.

[2] `https://www.privacy-regulation.eu/en/recital-26-GDPR.htm`

3.1 Corpus

We used the Wikipedia-based anonymization corpus introduced in [18][3]. To the best of our knowledge, this represents the largest manually annotated, domain-independent (*i.e.*, not medical or legal focused) corpus of manually anonymized documents. It comprises 553 biography abstracts randomly selected from a diverse set of public individuals, including athletes, politicians, and others. Because the sampling is random, many subjects are not widely known.

Manual annotations for document anonymization were performed following the privacy-oriented guidelines proposed in [20]. Five annotators labeled identifying terms and assigned semantic categories to them.

3.2 Anonymization methods

First, we considered traditional *named entity recognition (NER)-based methods*, which use neural models trained to detect predefined sets of *named entities* that are assumed to facilitate re-identification (*e.g.*, person, organization or location names). These include `Stanford NER` [12] with three models trained to detect different sets of NEs (`St.NER3`, `St.NER4` and `St.NER7`), the `spaCy` NE recognizer[4], and the Microsoft `Presidio` de-identification SDK[5].

Second, we evaluated two *PPDP-oriented approaches*, each in two configurations. The method in [7] employs Word2Vec embeddings to detect the most re-identifying terms according to their embedding-based similarity with the identity to be protected. For this, it uses a similarity threshold that we set to `Word2Vec t=0.5` and `Word2Vec t=0.25`, where a lower threshold results in stronger anonymization. The method in [18] enforces k-anonymity [22] on the documents by leveraging knowledge graphs through two different masking policies: k-`anonymity random` and k-`anonymity greedy`. We used $k = 5$ (as defined by the authors) for both methods, but `greedy` tends to mask fewer terms.

Third, we examined recent text anonymization methods based on *prompting LLMs* [16]: `Student de-identification` [23], Microsoft `Sparks` [2] and `Man vs the machine` [19]. All these methods employed the *gpt-4o-2024-09-03*[6] model. The first two approaches draw inspiration from NER, defining specific term types to be masked, while the latter adopts a more PPDP-oriented strategy, requesting the masking of any potentially disclosive term.

Finally, we also evaluated the `manual` anonymizations provided by the human annotators for the evaluation corpus [18].

3.3 Attack implementation

Our risk assessment framework was implemented considering a feasible threat scenario that emulates the resources and strategies that can realistically be expected of an attacker.

[3] `https://github.com/anthipapa/bootstrapping-anonymization`
[4] `https://spacy.io/api/entityrecognizer`
[5] `https://github.com/microsoft/presidio/releases/tag/2.2.354`
[6] `https://openai.com/index/hello-gpt-4o/`

This is reflected in the LLM that we used as reader: *Qwen3-4B* [24], specifically, the `Qwen3-4B-unsloth-bnb-4bit`[7] variant, which is a 4-bit quantized version that optimally balances performance with minimal (for an LLM) computational cost and memory allocation, requiring less than 16GB of VRAM. The rationale for choosing this model is that attackers are likely to prioritize efficiency in order to conduct large-scale automated attacks at minimal cost. As mentioned in Section 2.4, RAG compensates for the inherent limitations of smaller LLMs, ensuring robust performance despite the constraints of the model. On the other hand, the model's modest hardware requirements enable the attack to run on consumer-grade hardware, making it more feasible, attractive, and in line with GDPR's Recital 26.

We also used *thenlper/gte-small*[8] [10] as embedding model and the FAISS library[9] [4] for similarity retrieval. Both components were used as is, leveraging their readily available implementation. The retriever was set to obtain $k = 10$ textual chunks per document, with a maximum segment length of 128 tokens.

As BK, we used the bodies of the Wikipedia articles of the anonymized biographies' abstracts. Nevertheless, we found that approximately 8% of the biographies lacked a body section, which constrains the maximum re-identification risk to 92%. This defines a worst-case scenario (ideal for privacy evaluation) where the BK encompasses most of the information in the anonymized documents, but keeps it realistic, as it does not directly overlap with all anonymized documents or include knowledge for all individuals.

With this setting, processing the entire dataset required a total runtime of 1 hour and 4 minutes in our hardware configuration, with around 680 milliseconds allotted for assessing each anonymized document.

The final step to simulate a realistic threat involves verifying that the LLM does not possess enough knowledge of the protected documents to significantly bias the risk assessment. Because LLMs are pre-trained on extensive Web data, they may know public domain details about anyone, especially well-known individuals. This means that the model may already incorporate parts of the simulated BK, which consists of publicly accessible information.

Specifically, in our case, the original (non-anonymized) target documents—Wikipedia biographical abstracts—were likely included in the LLM's training data. If these biographies, even largely unpopular, left a significant trace in the LLM's knowledge, this could bias the results by artificially elevating re-identification risk compared to a standard data release in which the unprotected documents are not public.

To assess this potential bias, we measured the risk when applying our method without BK. We did this using a prompt that contained only the target document and a list of ten candidate names, setting the random guess level at 10%. The model must therefore rely solely on its pre-trained knowledge and any directly re-identifying information that may remain in the anonymized documents. If the

[7] `https://huggingface.co/unsloth/Qwen3-4B-unsloth-bnb-4bit`
[8] `https://huggingface.co/thenlper/gte-small`
[9] `https://faiss.ai/`

LLM fails to re-identify the documents (*i.e.*, achieving a success rate around the random guessing) with this ablated version, this would indicate the absence of bias.

3.4 Results

Table 1 reports the re-identification risk as measured by the standard recall metric (employing the TAE library[10]) and by our approach. Since recall evaluates protection rather than risk, we calculated 100% - Recall to facilitate direct comparison. Additionally, risks assessed with our method are evaluated with and without BK, as described in Section 3.3.

Table 1. Re-identification risk by recall and our approach, with and without background knowledge.

Method	100%-Recall	Ours wo/BK	Ours w/BK
St. NER3	54.29%	27.12%	61.48%
St. NER4	42.29%	15.37%	45.75%
St. NER7	45.29%	30.02%	64.38%
spaCy	16.34%	12.84%	39.96%
Presidio	42.53%	21.34%	55.88%
Word2Vec $t = 0.5$	27.24%	17.18%	35.80%
Word2Vec $t = 0.25$	12.02%	12.30%	29.48%
$k = 5$-anonymity random	12.10%	9.95%	32.91%
$k = 5$-anonymity greedy	12.82%	11.57%	36.71%
Student de-identification	43.21%	26.40%	52.44%
Sparks	13.31%	15.73%	34.18%
Man vs the machine	9.28%	8.32%	23.69%
Manual	0.00%	15.01%	44.12%

The risks associated with the ablated version of our approach closely approximate those of random guessing for most methods. The only exceptions were the less protective methods (mostly NE-based methods), which likely missed (parts of) the individual's name. This indicates that the model has no memorized prior knowledge of the protected documents and, thus, an absence of bias.

Under this premise, recall-based assessment consistently underestimated the risk values empirically determined by our approach with BK. Furthermore, recall often yielded misleading assessments. For instance, techniques with recall-based risk in the narrow range of 42% to 45% exhibited actual risks between 45% and 64% according to our framework.

[10] `https://github.com/BenetManzanaresSalor/TAE`

In line with this pattern, manual anonymization ranked the worst eighth in risk, highlighting that human-generated annotations, due to their subjectivity and susceptibility to errors, are unreliable for privacy assessment.

Regarding anonymization methods, NER-based approaches provided the weakest protection. In contrast, PPDP-oriented methods achieved notably lower risks. Meanwhile, LLM-based methods yielded varied results: the first one performs no better than NER-based techniques, the second one reaches parity with PPDP-oriented methods, and the last one (which employs a PPDP-oriented strategy) provides the lowest observed risk.

4 Conclusions and future work

We have introduced a novel privacy evaluation framework that leverages RAG and LLMs to automatically quantify the residual re-identification risk in anonymized documents. The approach simulates a feasible AI-powered adversary performing the textual equivalent of record linkage used in structured databases [1, 8]. This work establishes our metric as a practical foundation for benchmarking anonymization methods under realistic threat models.

Empirical experiments revealed that the risk measured by our approach consistently and substantially exceeded human-based assessments, exposing their limitations against AI-powered adversaries. Furthermore, results suggest that NER-based techniques are not competitive, as they systematically provided the least effective protection. This outcome aligns with their reliance on predefined entity types, which overlooks other terms that may facilitate re-identification [7]. In contrast, PPDP-oriented approaches, which are not constrained by entity types, achieved lower re-identification risks. LLM-based methods produced diverse results, underscoring the substantial influence of prompt-based configuration on risk levels. Nevertheless, they emerged as a promising alternative, occasionally providing the best protection while offering great practicality through their straightforward, prompt-driven methodology.

The substantial risk levels reported by our framework call for stronger protection standards in open environments such as the Web. This extends to human-based anonymization, which resulted in unexpectedly high risks, and underperformed compared to PPDP-oriented, most LLM-based and even a NER-based approach. This proves it as an unreliable ground truth for privacy assessment.

Future work includes assessing the impact of different (and larger) LLMs, retrieval, and embedding models. This will enable us to study the trade-off between the effectiveness of the attack and the resources required to carry it out. A key challenge is data contamination, as LLMs trained on vast web data may have already encountered the public corpora used for evaluation.

Acknowledgments. We acknowledge support from the Government of Catalonia (Acadèmia d'Excel.lència Prize to D. Sánchez), MCIN/AEI under grant PID2024-157271NB-I00 "CLEARING-IT" and the European Commission (project HORIZON-101292277 "SoBigDataIP"). B. Manzanares-Salor is also supported by the Spanish Government under an FPU grant (ref. FPU23/01785).

Bibliography

[1] Abril, D., Torra, V., Navarro-Arribas, G.: Supervised learning using a symmetric bilinear form for record linkage. Information Fusion **26**, 144–153 (2015)

[2] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S.M., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: Early experiments with GPT-4. CoRR **abs/2303.12712** (2023)

[3] Domingo-Ferrer, J., Sánchez, D., Soria-Comas, J.: Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections, vol. 8. Synthesis Lectures on Information Security, Privacy, and Trust (2016)

[4] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. IEEE Transactions on Big Data pp. 1–17 (2025)

[5] Dwork, C.: Differential privacy. In: International Colloquium on Automata, Languages and Programming. vol. 4052, pp. 1–12. Springer (2006)

[6] European Commission: Regulation (EU) 2016/679 (General Data Protection Regulation) (2016), `https://eur-lex.europa.eu/eli/reg/2016/679/oj`

[7] Hassan, F., Sánchez, D., Domingo-Ferrer, J.: Utility-preserving privacy protection of textual documents via word embeddings. IEEE Transactions on Knowledge and Data Engineering **35**, 1058–1071 (2023)

[8] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., De Wolf, P.P.: Statistical Disclosure Control, vol. 2. Wiley New York, New York (2012)

[9] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. pp. 9459–9474 (2020)

[10] Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M.: Towards general text embeddings with multi-stage contrastive learning. CoRR **abs/2308.03281** (2023)

[11] Lison, P., Pilán, I., Sánchez, D., Batet, M., Øvrelid, L.: Anonymisation models for text data: State of the art, challenges and future directions. In: 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 (2021)

[12] Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the ACL: System Demonstrations. pp. 55–60. Association for Computational Linguistics (2014)

[13] Manzanares-Salor, B., Sánchez, D., Lison, P.: Automatic evaluation of disclosure risks of text anonymization methods. In: Domingo-Ferrer, J., Laurent, M. (eds.) Privacy in Statistical Databases - International Conference, PSD 2022, Paris, France, September 21-23, 2022, Proceedings. Lecture Notes in Computer Science, vol. 13463, pp. 157–171. Springer (2022)
[14] Manzanares-Salor, B., Sánchez, D., Lison, P.: Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack. Data Min. Knowl. Discov. **38**(6), 4040–4075 (2024)
[15] Manzanares-Salor, B., Sánchez, D.: Enhancing text anonymization via re-identification risk-based explainability. Knowledge-Based Systems **310**, 112945 (2025)
[16] Manzanares-Salor, B., Sánchez, D.: A comparative analysis, enhancement and evaluation of text anonymization with pre-trained large language models. Expert Systems with Applications **297**, 129474 (2026)
[17] Manzanares-Salor, B., Sánchez, D., Lison, P.: Unsupervised utility evaluation of text anonymization methods via neural language models. Neural Networks **202**, 109079 (2026)
[18] Papadopoulou, A., Lison, P., Øvrelid, L., Pilán, I.: Bootstrapping text anonymization models with distant supervision. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 4477–4487. European Language Resources Association, Marseille, France (Jun 2022)
[19] Patsakis, C., Lykousas, N.: Man vs the machine in the struggle for effective text anonymisation in the age of large language models. Scientific Reports **13**(1), 16026 (2023)
[20] Pilán, I., Lison, P., Ørelid, L., Papadopoulou, A., Sánchez, D., Batet, M.: The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. Computational Linguistics **48**(8), 1053–1101 (2022)
[21] Pilán, I., Manzanares-Salor, B., Sánchez, D., Lison, P.: Truthful text sanitization guided by inference attacks. Applied Soft Computing **185**, 114013 (2025)
[22] Samarati, P.: Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering **13**(6), 1010–1027 (2001)
[23] Singhal, S., Zambrano, A.F., Pankiewicz, M., Liu, X., Porter, C., Baker, R.S.: De-identifying student personally identifying information with gpt-4. In: Proceedings of the 17th International Conference on Educational Data Mining. pp. 559–565 (2024)
[24] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. CoRR **abs/2505.09388** (2025)