A Comparative Study of Large Language Models and Traditional Privacy Measures to Evaluate Query Obfuscation Approaches

Francesco Luigi De Faveri*
University of Padova
Padova, Italy
francescoluigi.defaveri@phd.unipd.it

Guglielmo Faggioli
University of Padova
Padova, Italy
guglielmo.faggioli@unipd.it

Nicola Ferro
University of Padova
Padova, Italy
nicola.ferro@unipd.it

Abstract

When interacting with an Information Retrieval (IR) system, users might disclose personal information, such as medical details, through their queries. Thus, assessing the level of privacy granted to users when querying an IR system is essential to determine the confidentiality of submitted sensitive data. Query obfuscation protocols have traditionally been employed to obscure a user's real information need when retrieving documents. In these protocols, the query is modified employing ε-Differential Privacy (DP) obfuscation mechanisms, which alter query terms according to a predefined privacy budget ε. While this budget ensures formal mathematical guarantees, it provides only limited guarantees of the privacy experienced by the user and calls for empirical privacy evaluation to be carried out. Such privacy assessments employ lexical and semantic similarity measures between the original and obfuscated queries. In this study, we explore the role of Large Language Models (LLMs) in privacy evaluation, simulating a scenario where users employ such models to determine whether their input has been effectively privatized. Our primary research objective is to determine whether LLMs provide a novel perspective on privacy estimation and if their assessments serve as a proxy for traditional similarity metrics, such as the Jaccard and cosine similarity derived from Transformer-based sentence embeddings. Our findings reveal a positive correlation between LLMs-generated privacy scores and cosine similarity computed using different Transformer architectures. This suggests that LLM assessments act as a proxy for similarity-based measures.

CCS Concepts

• **Security and privacy** → **Privacy-preserving protocols**; **Privacy protections**; **Usability in security and privacy**.

Keywords

Privacy Evaluation, Query Processing, Large Language Models, Information Retrieval, Information Security

ACM Reference Format:
Francesco Luigi De Faveri, Guglielmo Faggioli, and Nicola Ferro. 2025. A Comparative Study of Large Language Models and Traditional Privacy Measures to Evaluate Query Obfuscation Approaches. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3726302.3730158

*Corresponding Author.

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730158

1 Introduction

Users' sensitive information is constantly shared with search engines, online social networks, and smart devices, often at the expense of privacy [1, 5, 7, 42]. For example, users often interrogate search engines with their medical information and symptoms [1]. Thus, ensuring appropriate privacy protections for users interacting with Information Retrieval (IR) systems is paramount. Privacy measures are a mandatory requirement for complying with privacy regulations [21], such as the General Data Protection Regulation (GDPR) [14], while also fostering the development of trustworthy intelligent systems and algorithms. ε-Differential Privacy (DP) [13] is deemed to be the gold standard to provide formal privacy guarantees during data processing. Broadly speaking, DP operates by injecting noise, whose strength is regulated by parameter ε, also known as *Privacy Budget*. However, assessing the privacy provided to the user query cannot rely only on the magnitude of ε as the effects of a DP mechanism on the final result—hence the actual privacy it provides—depends on several other aspects, such as the underlying distribution and the type of processing carried out [30, 34].

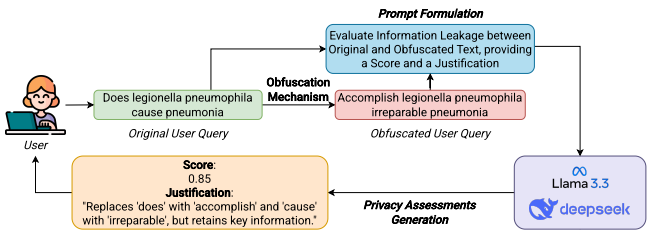

Figure 1: Example of employing LLMs to evaluate query obfuscations, justifying to users the assigned score.

In this study, we explore the use of Large Language Models (LLMs) as pseudo-assessors for evaluating privacy in query obfuscation protocols, as illustrated in Figure 1. Specifically, we propose the adoption of a LLM after the query has been obfuscated using an obfuscation mechanism to assess if the produced text retains identifiable traces of the original information need. The contributions of this work are structured as follows: i) we delineate the formal methodology of judging if a text has been obfuscated adequately by an obfuscation mechanism; ii) we propose the use of LLMs for assessing the privacy of queries obfuscated using different ε-DP approaches, mirroring a pseudo-relevance assessment of the privacy, and iii) we show the LLMs scores correlation with traditional

metrics used to determine the effectiveness of query obfuscations. Our findings show that, in general, LLM generated labels positively correlate with the traditional privacy assessment measures, such as lexical overlap and semantic similarity. In detail, we observe that the LLM-based labels encompass both semantic and lexical aspects, providing a different perspective respective to either approach.

Section 2 presents an overview of the query obfuscation protocol, explaining the traditional methods used to evaluate privacy. Section 3 details the methodology used for generating privacy assessments with LLMs to evaluate ε-DP texts. Section 4 explains the experimental setup, discussing the statistical findings between LLMs, used as privacy assessors, and standard privacy metrics. Finally, Section 5 summarizes the findings and outlines future work.

2 Query Obfuscation Protocol

Query obfuscation protocols [2, 16, 19] are a class of privacy preserving strategies used to protect user confidential information when interacting with IR systems. These protocols work under the assumption that the IR system is non-collaborative towards protecting user privacy, i.e., it does not implement any privacy mechanism to safeguard sensitive information needs. Hence, it represents an optimal tradeoff when the user desires to protect their privacy and the IR system does not implement any encrypted access to the information. The protocol works as follows: on the client side, which is considered safe, the text of the original query is transformed by an obfuscation mechanism, i.e., an algorithm that accepts the query text as input, masks the original information need, and outputs one or more obfuscated queries. The obfuscated queries are submitted to the IR system — considered unsafe —, which retrieves and ranks the documents in response to such queries. User privacy is protected since the IR system does not know the original query, and the ranked list corresponds to a modified query. Yet, if the obfuscation was effective, the ranked list returned to the user is expected to contain some documents relevant to the original query, possibly at low ranks. Therefore, the document set returned is re-ranked on the client's side using the original user query. This does not present privacy risks as only the client knows the original query. In this protocol, the user trades part of the retrieval performance for enhancing privacy, hiding their actual information need to the IR system that acts as the adversary during the documents' retrieval.

The ε-DP framework [13] provides a formal definition of privacy to the text, ensuring the confidentiality of the information contained in the texts by employing randomization during the query obfuscation phase. The level of formal privacy is controlled by the *Privacy Budget* parameter $\varepsilon \in [0, +\infty)$, which regulates the amount of statistical noise added query terms [18, 37, 38] or influences the sampling probabilities for generating obfuscated terms [3, 4, 6, 39]. Nevertheless, ε cannot be considered a perfect proxy of the actual privacy experienced by the user, as its effects depend on several other aspects, such as the underlying data distribution and the approach used to represent the text.

2.1 Privacy Measures

Assessing the privacy provided by an ε-DP mechanism remains a well-established challenge within the research community [30, 34]. Wagner and Eckhoff [34] define a set of aspects to assess the obfuscation capabilities of a mechanism: lexical similarity, semantic similarity and failure rates. *Lexical similarity* quantifies the term overlap between the original and the obfuscated texts. This metric is typically assessed using indicators such as the Jaccard Score, BLEU [24], and ROUGE [23]. On the other hand, the *semantic similarity* usually employs Transformers [33] or BERT Scores [26, 41]. Specifically, considering a Transformer $\mathcal{T}$, the semantic similarity between the original and obfuscated query text, respectively q and $\tilde{q}$, is computed as the cosine similarity cs between the query embeddings in the latent space, i.e., $cs = \frac{\mathcal{T}(q)\cdot\mathcal{T}(\tilde{q})}{\|\mathcal{T}(q)\|\|\mathcal{T}(\tilde{q})\|}$. *Failure rates* [22, 27], i.e., N_w and S_w, measure the probability of masking a word w with itself (N_w) and the size of the words that are used to mask the same term (S_w). Notice that, these measures are useful only for word-level obfuscation mechanisms and completely neglect the fact that a word can be obfuscated with a synonym.

3 Privacy Assessments Generation

In this section, we propose an LLM-based methodology to evaluate the effectiveness of an obfuscation mechanism. When determining if the query's text has been obfuscated, multiple aspects should be considered. For instance, minor modifications–such as altering a few characters in a term or changing the term order–can significantly change the text's overall meaning. Conversely, two sentences may differ syntactically while retaining the same semantic meaning. Traditional privacy evaluation metrics based on lexical similarity between original and obfuscated queries can be trivially fooled using synonyms to replace the query terms. Conversely, semantic similarity measured by transformers can be more robust towards identifying the similarities between texts. However, employing Transformers to determine whether a text has been obfuscated presents limitations when the text is rephrased. For example, encoding the sentences *"Mr. Doe was born in 1985 and lives in LA."* and *"John D., in his 40s, lives in Los Angeles."* using MiniLM [26] and computing the cosine similarity between their embeddings is 0.56, which poorly reflects the absence of privacy if we obfuscate the first sentence with the second. Despite reducing the cosine between the two texts, rephrasing a sentence does not ensure adequate privacy.

"Evaluate the information leakage from the original text to the obfuscated texts, providing a justification for each score given. Consider lexical and semantic similarities between original and obfuscated texts. The score should be an integer/float between **min** and **Max**, where **min** indicates no information leakage, and **Max** indicates complete information leakage.
The original text is: **original_text**. The obfuscated texts are: **obfuscated_texts**."

Figure 2: Prompt Template format submitted to the LLMs.

On a different research line, when it comes to IR evaluation, several studies [15, 32, 35, 40] investigated the possibility of using LLMs to judge relevance. However, to the best of our knowledge, no prior research has explored the application of LLMs for privacy assessments of textual data. To address this gap, we propose leveraging LLMs to assess privacy, providing the first experimental insights on the LLMs capabilities of understanding privacy, limiting assessment costs and time. In this task, both lexical and contextual aspects—traditionally considered in privacy relevance assessments [12, 36]—must be jointly analyzed to understand the extent

of information leakage from obfuscated versions of queries. Consequently, we develop a prompt to ask a LLM to evaluate the privacy levels attained by an obfuscated query compared to the original one, extending beyond conventional evaluation metrics. The general prompt template used in our experimental evaluation is presented in Figure 2. This template takes the **original_text** as the reference and the **obfuscated_texts** as a set of corresponding obfuscated versions. Additionally, it specifies the expected output score domain (integer or floating values) and the key aspects to consider when evaluating privacy, i.e., lexical and contextual similarity. The template also requires justification with each score assigned by the LLM, ensuring a comprehensive leakage assessment.

4 Experiments

4.1 Setup

To empirically test the proposed methodology, we consider two TREC collections MSMARCO Deep Learning 2019 track (DL'19)[8], consisting of 43 queries, and the TREC Medline 2004 collection (Med'04) [28], comprising 50 queries. Adopting the Med'04 queries represents a real obfuscation scenario, where the user is interested in finding information about a disease and, thus, aims to protect the confidentiality of the queries. We employ the pyPANTERA Python package[17], consider $\varepsilon \in \{1, 5, 10, 12.5, 15, 17.5, 20, 25, 30, 50\}$ and generate 10 obfuscated versions per query. We apply four state-of-the-art DP mechanisms implemented in the package, namely Cumulative Multivariate Perturbation (CMP) [18], Mahalanobis perturbation [37], and their respective Vickrey's variants [38]. Considering the benchmarks available on the Artificial Analysis platform[1], we selected the two highest-performing open-source LLMs, i.e., one reasoning-oriented model, DeepSeek-R1 [11]-distill-Llama70b a fine-tuned version of Llama 3.3 70B using samples generated by DeepSeek-R1, and the standard version of LLama 3.3 70B [31]. The prompt forces the LLM response to be in JSON format. The code, the results and the appendix are available in the repository[2].

4.2 Key Findings

4.2.1 Changing the prompt: Continuous and Discrete Privacy Scores. We test two different prompts for obtaining the privacy assessments. The LLMs are asked to provide: i) an information leakage score in a continuous interval, i.e., ranging in the [0,1] interval where 0 means no information leakage from the private query, and 1 means total information leakage; ii) a discrete value using a score in a Likert scale [9, 29] from a minimum score of 1, indicating that no information is understandable from the obfuscated query to a maximum score of 5, suggesting that the obfuscated query is identical to the original text. The prompts adopted for getting the scores from the LLMs are available in the paper's online appendix.

To avoid encumbering, we report only the results on the Med'04 queries of two mechanisms (Mahalanobis and VickreyMhl) for three privacy setups, $\varepsilon \in \{1, 15, 50\}$. The results on DL'19 and other mechanisms are equivalent and available in the paper repository.

Figure 3 presents the score distributions for the Continuous and Discrete prompts employed to evaluate different obfuscation mechanisms. The results indicate that the distributions of scores exhibit

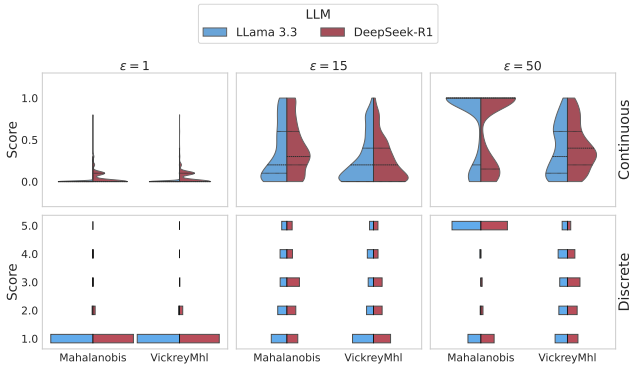

Figure 3: Changing the Prompts on the Med'04 queries. The Continuous score distributions report also the quartiles.

similar patterns across the different prompting strategies used to obtain the privacy assessments. Under a strong privacy regime, i.e., $\varepsilon = 1$, the LLMs consistently evaluate queries as highly obfuscated for both mechanisms, yielding a low information leakage score centring the distribution around 0.0 for continuous scores while frequently assigning a score of 1 for the discrete prompt. DeepSeek-R1 identifies more information leakage compared to LLama 3.3. After manual inspection, the cases where DeepSeek-R1 identifies more leakage appear overestimated, suggesting that DeepSeek-R1 is particularly conservative and should be favoured when the user wishes to attain strong privacy guarantees. As ε increases, the degree of obfuscation applied to the textual data decreases, leading to a shift in privacy assessments toward higher information scores for DeepSeek-R1, with most privacy scores around 0.3. At $\varepsilon = 15$, a distinction emerges in the obfuscation strategies, as the VickreyMhl mechanism tends to have lower leakage values: this is in line with previous research [10, 16, 17], indicating that the evaluation approach is consistent. Finally, at $\varepsilon = 50$, Mahalanobis often fails in obfuscating the query, as testified by the large number of obfuscation queries labelled with 1 in the continuous case or 5 in the discrete case. VickreyMhl provides a more satisfactory degree of privacy, with its score distribution centred around 0.5, demonstrating the same conclusions found in [17, 38] with standard measures.

4.2.2 LLMs Privacy Scores & Traditional Privacy Analysis. This section compares the privacy scores obtained from the LLMs using the prompt that generated a score in the [0,1] range. As traditional privacy measures to which we compare the LLMs score, we employ three Transformers [33] architecture, namely MiniLM [25], DistilRoBERTa [26], and MPNET [20], from the Huggingface platform[3], to compute the cosine similarity between query obfuscations, and lexical metrics using the Jaccard, BLEU, and ROUGE scores, cf. Section 2.1. To validate the correctness of the LLMs used to assess privacy, we compute the correlation between the LLMs scores and the cosine similarities of the transformers measured as Kendall's, Pearson's and Spearman's correlations. On the other hand, we measure the Mean Squared Error between lexical metrics and the LLMs privacy assessments across different privacy budgets ε.

[1] https://artificialanalysis.ai/
[2] https://github.com/Kekkodf/LLM4PrivacyEval
[3] https://huggingface.co/

Table 1: Correlation statistics of LLMs scores and Cosine Similarity obtained considering the aggregation of the ε parameters confronting different Transformers and Mechanisms. The results are organized by obfuscation mechanism and collections.

	LLM	Transformer	CMP			Mahalanobis			VickreyCMP			VickreyMhl		
			Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman
DL'19	*DeepSeek-R1*	*MiniLM*	0.742	0.872	0.889	0.702	0.864	0.854	0.574	0.744	0.743	0.594	0.775	0.762
		DistilRoBERTa	0.739	0.876	0.886	0.702	0.868	0.853	0.569	0.740	0.738	0.591	0.776	0.759
		MPNET	0.735	0.869	0.884	0.700	0.859	0.853	0.579	0.747	0.750	0.592	0.773	0.761
	LLama 3.3	*MiniLM*	0.802	0.924	0.929	0.765	0.905	0.901	0.644	0.801	0.806	0.645	0.812	0.806
		DistilRoBERTa	0.793	0.920	0.924	0.765	0.905	0.901	0.636	0.790	0.798	0.634	0.807	0.793
		MPNET	0.792	0.919	0.924	0.761	0.902	0.900	0.650	0.803	0.811	0.641	0.813	0.804
Med'04	*DeepSeek-R1*	*MiniLM*	0.721	0.875	0.869	0.691	0.866	0.841	0.576	0.753	0.748	0.563	0.762	0.729
		DistilRoBERTa	0.711	0.870	0.862	0.683	0.865	0.835	0.577	0.752	0.747	0.558	0.759	0.723
		MPNET	0.717	0.878	0.868	0.687	0.867	0.839	0.562	0.742	0.735	0.552	0.753	0.716
	LLama 3.3	*MiniLM*	0.754	0.883	0.879	0.715	0.853	0.850	0.620	0.778	0.785	0.593	0.767	0.752
		DistilRoBERTa	0.747	0.881	0.874	0.711	0.854	0.848	0.609	0.771	0.774	0.581	0.755	0.741
		MPNET	0.752	0.886	0.881	0.708	0.854	0.844	0.606	0.769	0.770	0.583	0.759	0.741

Table 2: Mean Squared Errors between traditional privacy evaluation measures, i.e., BLEU and ROUGE, and the LLM scores in the [0,1] range across different ε-Privacy Budgets on the obfuscated queries in the DL'19 and Med'04 collections.

			DL'19										Med'04									
	LLM	Mechansim	1.0	5.0	10.0	12.5	15.0	17.5	20.0	25.0	30.0	50.0	1.0	5.0	10.0	12.5	15.0	17.5	20.0	25.0	30.0	50.0
BLEU	*DeepSeek-R1*	Mahalanobis	0.008	0.009	0.057	0.074	0.090	0.074	0.088	0.059	0.060	0.072	0.007	0.014	0.070	0.098	0.138	0.122	0.130	0.124	0.102	0.105
		VickreyMhl	0.007	0.010	0.037	0.059	0.085	0.110	0.114	0.118	0.134	0.186	0.008	0.010	0.061	0.08	0.103	0.122	0.125	0.156	0.157	0.183
	LLama 3.3	Mahalanobis	0.003	0.002	0.061	0.130	0.164	0.161	0.139	0.082	0.050	0.045	0.001	0.009	0.067	0.097	0.121	0.124	0.171	0.155	0.093	0.087
		VickreyMhl	0.001	0.005	0.030	0.067	0.105	0.143	0.174	0.207	0.219	0.255	0.001	0.004	0.040	0.085	0.110	0.116	0.124	0.123	0.134	0.172
ROUGE	*DeepSeek-R1*	Mahalanobis	0.007	0.009	0.030	0.030	0.044	0.055	0.113	0.095	0.097	0.145	0.007	0.013	0.033	0.028	0.043	0.069	0.105	0.084	0.103	0.089
		VickreyMhl	0.007	0.009	0.024	0.030	0.041	0.043	0.043	0.049	0.050	0.064	0.008	0.010	0.039	0.038	0.039	0.044	0.039	0.048	0.052	0.065
	LLama 3.3	Mahalanobis	0.003	0.002	0.036	0.055	0.043	0.037	0.050	0.052	0.050	0.065	0.001	0.008	0.031	0.035	0.049	0.091	0.129	0.148	0.070	0.066
		VickreyMhl	0.001	0.005	0.018	0.034	0.048	0.058	0.064	0.068	0.066	0.072	0.001	0.004	0.023	0.039	0.045	0.040	0.052	0.049	0.052	0.063

Table 1 presents the correlation, organized by obfuscation mechanism, between the LLMs privacy assessment scores and the cosine similarities calculated by the three different transformer models on the obfuscated queries in the DL'19 and Med'04. Our findings show that for all the mechanisms tested, Kendall's correlation is strongly positive and non-pathological, i.e., equal to 1, showing that while the measures agree on assessing the privacy computed, they consider different aspects. The correlation decreases when evaluating Vickrey's variants of the CMP and Mahalanobis mechanisms, yet strong positive correlations between the measures are retained.

On the other hand, Table 2 reports the Mean Squared Errors between BLEU and ROUGE scores with the LLMs assessed privacy scores for Mahalanobis and its Vicrey variant. The complete table of the results, considering also CMP and Vickrey CMP as mechanisms, and the Jaccard Index as privacy measure, is available in the repository and shows comparable patterns to the ones reported in Table 2. For low ε values, i.e., $\varepsilon \in \{1, 5\}$, the errors obtained by confronting all traditional scores against the LLMs scores are consistently below the 10%, either for DeepSeek-R1 and LLama 3.3. When increasing the privacy budget ε, the BLEU score is the most different from the LLMs's ones, with errors rising to a maximum of 25.5% for LLama 3.3 and VickreyMhl. Finally, the results show a strong positive correlation with contextual similarity and minor errors with lexical metrics. Therefore, LLMs generated assessments can capture semantic and lexical facets into a unique privacy score.

5 Conclusion

In this study, we considered the problem of measuring privacy when user queries are obfuscated employing ε-DP mechanisms. Standard evaluation measures consider the lexical and contextual similarities between original and obfuscated texts. We test the use of LLMs as privacy assessors to determine if the obfuscated query leaks information from the original query. We empirically show that the scores generated by the LLMs combine the effectiveness of traditional lexical measures and semantic similarity-based approaches used in privacy assessments. Our findings indicate that LLMs generates information leakage scores using continuous and discrete Likert-like annotations. In addition, by computing the correlation between traditional and LLMs-based leakage scores, we observed a positive correlation with traditional semantic measures, suggesting that the method is consistent with past findings and minor differences compared to lexical measures. Consequently, LLMs-based leakage assessment represents a practical tradeoff between lexical and semantic privacy measures currently used to evaluate obfuscation mechanisms. Future work will combine human evaluation to assess whether the LLMs scores are accurately given from a human perspective, and the correlation of these LLMs generated scores with other actual evaluations of privacy, like the ones proposed in [10]. Finally, we plan to analyse which activation patterns are employed by LLMs during privacy assessments, bridging the gap between privacy generation scores and explainability of the models.

References

[1] Pol Mac Aonghusa and Douglas J. Leith. 2016. Don't Let Google Know I'm Lonely. *ACM Trans. Priv. Secur.* 19, 1 (2016), 3:1–3:25. doi:10.1145/2937754

[2] Danushka Bollegala, Tomoya Machide, and Ken-ichi Kawarabayashi. 2022. Query Obfuscation by Semantic Decomposition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, 6200–6211. https://aclanthology.org/2022.lrec-1.667

[3] Danushka Bollegala, Shuichi Otake, Tomoya Machide, and Ken-ichi Kawarabayashi. 2024. A Metric Differential Privacy Mechanism for Sentence Embeddings. *ACM Trans. Priv. Secur.* (Dec. 2024). doi:10.1145/3708321 Just Accepted.

[4] Ricardo Silva Carvalho, Theodore Vasiloudis, Oluwaseyi Feyisetan, and Ke Wang. 2023. TEM: High Utility Metric Differential Privacy on Text. In *Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023*, Shashi Shekhar, Zhi-Hua Zhou, Yao-Yi Chiang, and Gregor Stiglic (Eds.). SIAM, 883–890. doi:10.1137/1.9781611977653.CH99

[5] George Chalhoub and Ivan Flechais. 2020. "Alexa, Are You Spying on Me?": Exploring the Effect of User Experience on the Security and Privacy of Smart Speaker Users. In *HCI for Cybersecurity, Privacy and Trust - Second International Conference, HCI-CPT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12210)*, Abbas Moallem (Ed.). Springer, 305–325. doi:10.1007/978-3-030-50309-3_21

[6] Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A Customized Text Sanitization Mechanism with Differential Privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5747–5758. doi:10.18653/v1/2023.findings-acl.355

[7] Jonathan Cohn. 2016. My TiVo Thinks I'm Gay: Algorithmic Culture and Its Discontents. *Television & New Media* 17, 8 (2016), 675–690. doi:10.1177/1527476416644978 arXiv:https://doi.org/10.1177/1527476416644978

[8] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *CoRR* abs/2003.07820 (2020). arXiv:2003.07820 https://arxiv.org/abs/2003.07820

[9] Hugh M Culbertson. 1968. What is an Attitude? *The Journal of Extension* 6, 2 (1968), 9.

[10] Francesco Luigi De Faveri, Guglielmo Faggioli, and Nicola Ferro. 2025. Measuring Actual Privacy of Obfuscated Queries in Information Retrieval. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part I* (Lucca, Italy). Springer-Verlag, Berlin, Heidelberg, 49–66. doi:10.1007/978-3-031-88708-6_4

[11] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

[12] Tom Diethe, Oluwaseyi Feyisetan, Borja Balle, and Thomas Drake. 2020. Preserving privacy in analyses of textual data. (2020). https://www.amazon.science/publications/preserving-privacy-in-analyses-of-textual-data

[13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, Shai Halevi and Tal Rabin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 265–284.

[14] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. https://data.europa.eu/eli/reg/2016/679/oj

[15] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023*, Masaharu Yoshioka, Julia Kiseleva, and Mohammad Aliannejadi (Eds.). ACM, 39–50. doi:10.1145/3578337.3605136

[16] Guglielmo Faggioli and Nicola Ferro. 2024. Query Obfuscation for Information Retrieval Through Differential Privacy. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 14608)*, Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (Eds.). Springer, 278–294. doi:10.1007/978-3-031-56027-9_17

[17] Francesco Luigi De Faveri, Guglielmo Faggioli, and Nicola Ferro. 2024. pyPANTERA: A Python PAckage for Natural language obfuscaTion Enforcing pRivacy & Anonymization. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, Edoardo Serra and Francesca Spezzano (Eds.). ACM, 5348–5353. doi:10.1145/3627673.3679173

[18] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy- and Utility-Preserving Textual Analysis via Calibrated Multivariate Perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 178–186. doi:10.1145/3336191.3371856

[19] Arthur Gervais, Reza Shokri, Adish Singla, Srdjan Capkun, and Vincent Lenders. 2014. Quantifying Web-Search Privacy. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*, Gail-Joon Ahn, Moti Yung, and Ninghui Li (Eds.). ACM, 966–977. doi:10.1145/2660267.2660367

[20] Sai Muralidhar Jayanthi, Varsha Embar, and Karthik Raghunathan. 2021. Evaluating Pretrained Transformer Models for Entity Linking in Task-Oriented Dialog. *CoRR* abs/2112.08327 (2021). arXiv:2112.08327 https://arxiv.org/abs/2112.08327

[21] Alexandra Klymenko, Stephen Meisenbacher, Ali Asaf Polat, and Florian Matthes. 2025. A Systematic Analysis of Data Protection Regulations. (2025). doi:10.24251/HICSS.2025.535

[22] Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential Privacy in Natural Language Processing The Story So Far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, Oluwaseyi Feyisetan, Sepideh Ghanavati, Patricia Thaine, Ivan Habernal, and Fatemehsadat Mireshghallah (Eds.). Association for Computational Linguistics, Seattle, United States, 1–11. doi:10.18653/v1/2022.privatenlp-1.1

[23] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013/

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. doi:10.3115/1073083.1073135

[25] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. doi:10.18653/V1/D19-1410

[26] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 4512–4525. doi:10.18653/V1/2020.EMNLP-MAIN.365

[27] Alfréd Rényi. 1961. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, Vol. 4. University of California Press, 547–562.

[28] Patrick Ruch, Christine Chichester, Gilles Cohen, Frédéric Ehrler, Paul Fabry, Johan Marty, Henning Müller, and Antoine Geissbühler. 2004. Report on the TREC 2004 Experiment: Genomics Track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004 (NIST Special Publication, Vol. 500-261)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec13/papers/uhosp-geneva.geo.pdf

[29] Norbert Schwarz and Gerd Bohner. 2001. The construction of attitudes. *Blackwell handbook of social psychology: Intraindividual processes* (2001), 436–457.

[30] Samuel Sousa and Roman Kern. 2023. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. *Artif. Intell. Rev.* 56, 2 (2023), 1427–1492. doi:10.1007/S10462-022-10204-6

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023). doi:10.48550/ARXIV.2302.13971 arXiv:2302.13971

[32] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. *CoRR* abs/2406.06519 (2024). doi:10.48550/ARXIV.2406.06519 arXiv:2406.06519

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[34] Isabel Wagner and David Eckhoff. 2018. Technical Privacy Metrics: A Systematic Survey. *ACM Comput. Surv.* 51, 3 (2018), 57:1–57:38. doi:10.1145/3168389

[35] Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Quanquan Gu, Haifeng Chen, Wei Wang, and Wei Cheng. 2024. Large Language Models Can Be Contextual Privacy Protection Learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 14179–14201. https://aclanthology.org/2024.emnlp-main.785

[36] Rui Xin, Niloofar Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. 2024. A False Sense of Privacy: Evaluating Textual Data Sanitization Beyond Surface-level Privacy Leakage. In *Neurips Safe Generative AI Workshop 2024*. https://openreview.net/pdf?id=3JLtuCozOU

[37] Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020. A Differentially Private Text Perturbation Method Using Regularized Mahalanobis Metric. In *Proceedings of the Second Workshop on Privacy in NLP*. Association for Computational Linguistics. doi:10.18653/v1/2020.privatenlp-1.2

[38] Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2021. On a Utilitarian Approach to Privacy Preserving Text Generation. *CoRR* abs/2104.11838 (April 2021). doi:10.48550/ARXIV.2104.11838 arXiv:2104.11838 [cs.CL]

[39] Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential Privacy for Text Analytics via Natural Text Sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 3853–3866. doi:10.18653/v1/2021.findings-acl.337

[40] Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Are Large Language Models Good at Utility Judgments?. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 1941–1951. doi:10.1145/3626772.3657784

[41] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr

[42] Steven Zimmerman, Alistair Thorpe, Chris Fox, and Udo Kruschwitz. 2019. Privacy Nudging in Search: Investigating Potential Impacts. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019, Glasgow, Scotland, UK, March 10-14, 2019*, Leif Azzopardi, Martin Halvey, Ian Ruthven, Hideo Joho, Vanessa Murdock, and Pernilla Qvarfordt (Eds.). ACM, 283–287. doi:10.1145/3295750.3298952