Comparing LLMs and Traditional Privacy Measures to Evaluate Query Obfuscation Approaches*

Francesco Luigi, De Faveri[1], Guglielmo, Faggioli[1] and Nicola, Ferro[1]

[1]*Department of Information Engineering, University of Padova, Padova, Italy*

Abstract

When interacting with Information Retrieval (IR) systems, users may inadvertently disclose sensitive personal information, such as medical conditions, through their search queries. Therefore, evaluating the privacy these queries afford during the retrieval process is critical. Query obfuscation has traditionally been employed to hide a user's information need by modifying the original query. ε-Differential Privacy (ε-DP) mechanisms perturb query terms by a privacy budget ε and guarantee a grounded in mathematical proofs of privacy. However, ε does not fully capture the user's subjective experience of privacy, needing additional empirical tests. Privacy assessments typically rely on lexical and semantic similarity measures to quantify the difference between the original and obfuscated queries. In this work, we investigate how Large Language Models (LLMs) can be used to perform a privacy evaluation in such scenarios. Our central research question is whether LLMs can offer a new lens through which privacy can be assessed, and whether their scores correlate with similarity-based metrics, such as Jaccard similarity and cosine similarity between text embeddings. Our experimental results show a strong positive correlation between LLM-derived privacy assessments and cosine similarity values computed with different Transformers. These findings suggest that LLMs can effectively serve as proxies for traditional similarity measures in the context of privacy evaluation.

Keywords

Privacy Preserving Information Retrieval, Differential Privacy, Information Security, Large Language Models

1. Introduction

Users frequently disclose sensitive information—such as medical symptoms—when interacting with search engines, social platforms, or smart devices, often compromising their privacy [2, 3, 4, 5]. Ensuring adequate privacy protections in Information Retrieval (IR) systems is therefore essential to comply with regulations like the GDPR [6, 7]. ε-Differential Privacy (DP) [8] offers formal guarantees by injecting noise into data, with the privacy budget ε controlling the trade-off between utility and protection.

However, the actual privacy experienced by users depends not only on ε but also on other factors such as data distribution and processing characteristics [9, 10], making empirical evaluation necessary. In this work, we investigate the use of Large Language Models (LLMs) as pseudo-assessors to evaluate

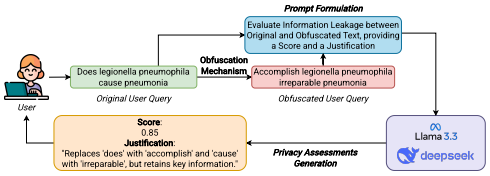

Figure 1: Example of employing LLMs to evaluate query obfuscations, justifying to users the assigned score.

the effectiveness of query obfuscation mechanisms (Figure 1). After obfuscating a query using an ε-DP mechanism, we prompt an LLM to judge whether the resulting text still reveals the original need. Our

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy

*This is an extended abstract of [1].

francescoluigi.defaveri@phd.unipd.it (F. L. De Faveri); guglielmo.faggioli@unipd.it (G. Faggioli); nicola.ferro@unipd.it (N. Ferro)

https://www.dei.unipd.it/~defaverifr/ (F. L. De Faveri); https://www.dei.unipd.it/~faggioli/ (G. Faggioli); https://www.dei.unipd.it/~ferro/ (N. Ferro)

0009-0005-8968-9485 (F. L. De Faveri); 0000-0002-5070-2049 (G. Faggioli); 0000-0001-9219-6239 (N. Ferro)

contributions are threefold: (i) we formalise the task of evaluating actual textual obfuscation, (ii) we propose LLM-based privacy evaluation as a proxy to traditional privacy metrics, and (iii) we show that LLM-generated scores correlate with established measures such as lexical and semantic similarity. LLM scores tend to integrate semantic and lexical aspects, offering a broader view of privacy evaluation.

2. Background

Query Obfuscation Protocol. Query obfuscation protocols [11, 12, 13] are a class of privacy preserving strategies used to protect user 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. On the client side, 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, unsafe, and retrieve and rank the documents in response to such queries. The ε-DP framework [8] provides a formal definition of privacy to the text, ensuring the privacy of the texts by employing randomisation during the query obfuscation phase. The level of formal privacy is controlled by the *Privacy Budget* parameter $\varepsilon \in \mathbb{R}_+$, which regulates the amount of statistical noise added query terms [14, 15, 16] or influences the sampling probabilities for generating obfuscated terms [17, 18, 19, 20]. ε cannot be considered a perfect proxy of the actual privacy experienced [21], as its effects depend on several aspects.

Privacy Measures. Assessing the privacy provided by an ε-DP mechanism remains a well-established challenge within the research community [9, 10]. Wagner and Eckhoff [9] define a set of aspects to assess the obfuscations 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 [22], and ROUGE [23]. On the other hand, the *semantic similarity* usually employs Transformers [24] or BERT Scores [25, 26]. Specifically, considering a Transformer $\mathcal{T}$, the semantic similarity between the original and obfuscated text, respectively q and $\tilde{q}$, is computed as the cosine similarity cs between the 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* [27, 28], 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

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 alter 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.

Prompt Template

"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**."

On a different research line, when it comes to IR evaluation, several studies [29, 30, 31, 32] 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 into the LLMs 's capabilities for understanding privacy, limiting assessment costs and time. In this task, both lexical and contextual aspects—traditionally considered in privacy relevance assessments [33, 34]—must be jointly analysed to understand the extent of information leakage from obfuscated versions of queries. Thus, 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. 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 wide leakage assessment.

4. Experimental Evaluation

To empirically test the methodology[1], we consider two TREC collections MSMARCO Deep Learning 2019 track (DL'19)[35], consisting of 43 queries, and the TREC Medline 2004 collection (Med'04) [36], 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[37] applying four state-of-the-art DP mechanisms implemented in it, namely Cumulative Multivariate Perturbation (CMP) [16], Mahalanobis perturbation [14], and their respective Vickrey's variants [15]. We selected two open-source LLMs, i.e., one reasoning-oriented model, DeepSeek-R1 [38]-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 [39].

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 [40, 41] 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 2 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 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. As ε increases, the degree of obfuscation applied to the textual data decreases, leading to a shift in privacy assessments toward higher information

[1]The code, the results and the appendix are available in the repository https://github.com/Kekkodf/LLM4PrivacyEval

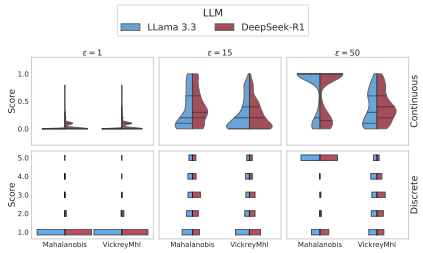

Figure 2: Changing the Prompts results. The Continuous score distributions report also the quartiles.

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 [13, 37, 21], indicating that the evaluation approach is consistent. Finally, at $\varepsilon = 50$, Mahalanobis often fails to obfuscate the query, as evidenced 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, demonstrating the same conclusions found in [15, 37] with standard measures.

Table 1
Correlation statistics of LLMs scores and Cosine Similarity obtained.

	LLM	Transformer	CMP			Mahalanobis			VickreyCMP			VickreyMhl		
			Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman
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

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 [24] architecture, namely MiniLM [42], DistilRoBERTa [26], and MPNET [43], to compute the cosine similarity between query obfuscations. Results on the lexical analysis are reported in the repository in the full paper [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. Table 1 presents the correlation, organised 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 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, yet strong positive correlations between the measures are retained.

5. Conclusion

This study addressed the challenge of evaluating privacy in user queries obfuscated through ε-DP mechanisms. While traditional approaches rely on lexical and semantic similarity between the original and obfuscated queries, we explored the use of LLMs as automated privacy assessors. Our empirical analysis shows that LLM-generated leakage scores effectively capture aspects of both lexical and semantic similarity, producing continuous and Likert-style outputs. The positive correlation with established semantic metrics indicates alignment with existing evaluation methods, while minor differences with lexical measures suggest a complementary perspective. Thus, LLM-based assessments offer a practical middle ground between current privacy evaluation techniques. Future work will include human judgments to validate LLM-based scores. Additionally, we aim to investigate the internal activation patterns of LLMs during privacy assessments, contributing to the trustworthiness of their evaluations.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for Readability and Spelling checks. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] F. L. De Faveri, G. Faggioli, N. Ferro, A comparative study of large language models and traditional privacy measures to evaluate query obfuscation approaches, in: N. Ferro, M. Maistro, G. Pasi, O. Alonso, A. Trotman, S. Verberne (Eds.), Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025, ACM, 2025, pp. 2711–2716. URL: https://doi.org/10.1145/3726302.3730158. doi:10.1145/3726302.3730158.

[2] J. Cohn, My tivo thinks i'm gay: Algorithmic culture and its discontents, Television & New Media 17 (2016) 675–690. URL: https://doi.org/10.1177/1527476416644978. doi:10.1177/1527476416644978. arXiv:https://doi.org/10.1177/1527476416644978.

[3] P. M. Aonghusa, D. J. Leith, Don't let google know i'm lonely, ACM Trans. Priv. Secur. 19 (2016) 3:1–3:25. URL: https://doi.org/10.1145/2937754. doi:10.1145/2937754.

[4] S. Zimmerman, A. Thorpe, C. Fox, U. Kruschwitz, Privacy nudging in search: Investigating potential impacts, in: L. Azzopardi, M. Halvey, I. Ruthven, H. Joho, V. Murdock, P. Qvarfordt (Eds.), Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019, Glasgow, Scotland, UK, March 10-14, 2019, ACM, 2019, pp. 283–287. URL: https://doi.org/10.1145/3295750.3298952. doi:10.1145/3295750.3298952.

[5] G. Chalhoub, I. Flechais, "alexa, are you spying on me?": Exploring the effect of user experience on the security and privacy of smart speaker users, in: A. Moallem (Ed.), 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, volume 12210 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 305–325. URL: https://doi.org/10.1007/978-3-030-50309-3_21. doi:10.1007/978-3-030-50309-3_21.

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

[7] A. Klymenko, S. Meisenbacher, A. A. Polat, F. Matthes, A systematic analysis of data protection regulations (2025). URL: https://hdl.handle.net/10125/109381. doi:10.24251/HICSS.2025.535.

[8] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: S. Halevi, T. Rabin (Eds.), Theory of Cryptography, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 265–284.

[9] I. Wagner, D. Eckhoff, Technical privacy metrics: A systematic survey, ACM Comput. Surv. 51 (2018) 57:1–57:38. URL: https://doi.org/10.1145/3168389. doi:10.1145/3168389.

[10] S. Sousa, R. Kern, How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing, Artif. Intell. Rev. 56 (2023) 1427–1492. URL: https://doi.org/10.1007/s10462-022-10204-6. doi:10.1007/S10462-022-10204-6.

[11] A. Gervais, R. Shokri, A. Singla, S. Capkun, V. Lenders, Quantifying web-search privacy, in: G. Ahn, M. Yung, N. Li (Eds.), Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014, ACM, 2014, pp. 966–977. URL: https://doi.org/10.1145/2660267.2660367. doi:10.1145/2660267.2660367.

[12] D. Bollegala, T. Machide, K. Kawarabayashi, Query obfuscation by semantic decomposition, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022, European Language Resources Association, 2022, pp. 6200–6211. URL: https://aclanthology.org/2022.lrec-1.667.

[13] G. Faggioli, N. Ferro, Query obfuscation for information retrieval through differential privacy, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part I, volume 14608 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 278–294. URL: https://doi.org/10.1007/978-3-031-56027-9_17. doi:10.1007/978-3-031-56027-9_17.

[14] Z. Xu, A. Aggarwal, O. Feyisetan, N. Teissier, A differentially private text perturbation method using regularized mahalanobis metric, in: Proceedings of the Second Workshop on Privacy in NLP, Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.privatenlp-1.2.

[15] Z. Xu, A. Aggarwal, O. Feyisetan, N. Teissier, On a utilitarian approach to privacy preserving text generation, CoRR abs/2104.11838 (2021). doi:10.48550/ARXIV.2104.11838. arXiv:2104.11838.

[16] O. Feyisetan, B. Balle, T. Drake, T. Diethe, Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations, in: J. Caverlee, X. B. Hu, M. Lalmas, W. Wang (Eds.), Proceedings of the 13th International Conference on Web Search and Data Mining, ACM, 2020, pp. 178–186. doi:10.1145/3336191.3371856.

[17] S. Chen, F. Mo, Y. Wang, C. Chen, J.-Y. Nie, C. Wang, J. Cui, A customized text sanitization mechanism with differential privacy, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5747–5758. URL: https://aclanthology.org/2023.findings-acl.355. doi:10.18653/v1/2023.findings-acl.355.

[18] R. S. Carvalho, T. Vasiloudis, O. Feyisetan, K. Wang, TEM: high utility metric differential privacy on text, in: S. Shekhar, Z. Zhou, Y. Chiang, G. Stiglic (Eds.), Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023, SIAM, 2023, pp. 883–890. URL: https://doi.org/10.1137/1.9781611977653.ch99. doi:10.1137/1.9781611977653.CH99.

[19] D. Bollegala, S. Otake, T. Machide, K.-i. Kawarabayashi, A metric differential privacy mechanism for sentence embeddings, ACM Trans. Priv. Secur. (2024). URL: https://doi.org/10.1145/3708321. doi:10.1145/3708321, just Accepted.

[20] X. Yue, M. Du, T. Wang, Y. Li, H. Sun, S. S. M. Chow, Differential privacy for text analytics via natural text sanitization, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 3853–3866. URL: https://aclanthology.org/2021.findings-acl.337. doi:10.18653/v1/2021.findings-acl.337.

[21] F. L. De Faveri, G. Faggioli, N. Ferro, 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, Springer-Verlag, Berlin, Heidelberg, 2025, p. 49–66. URL: https://doi.org/10.1007/978-3-031-88708-6_4. doi:10.1007/978-3-031-88708-6_4.

[22] K. Papineni, S. Roukos, T. Ward, W. Zhu, 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, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.

[23] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text

generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[26] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 4512–4525. URL: https://doi.org/10.18653/v1/2020.emnlp-main.365. doi:10.18653/V1/2020.EMNLP-MAIN.365.

[27] O. Klymenko, S. Meisenbacher, F. Matthes, Differential privacy in natural language processing the story so far, in: O. Feyisetan, S. Ghanavati, P. Thaine, I. Habernal, F. Mireshghallah (Eds.), Proceedings of the Fourth Workshop on Privacy in Natural Language Processing, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1–11. URL: https://aclanthology.org/2022.privatenlp-1.1. doi:10.18653/v1/2022.privatenlp-1.1.

[28] A. Rényi, 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, volume 4, University of California Press, 1961, pp. 547–562.

[29] G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, H. Wachsmuth, Perspectives on large language models for relevance judgment, in: M. Yoshioka, J. Kiseleva, M. Aliannejadi (Eds.), Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023, ACM, 2023, pp. 39–50. URL: https://doi.org/10.1145/3578337.3605136. doi:10.1145/3578337.3605136.

[30] S. Upadhyay, R. Pradeep, N. Thakur, N. Craswell, J. Lin, UMBRELA: umbrela is the (open-source reproduction of the) bing relevance assessor, CoRR abs/2406.06519 (2024). URL: https://doi.org/10.48550/arXiv.2406.06519. doi:10.48550/ARXIV.2406.06519. arXiv:2406.06519.

[31] H. Zhang, R. Zhang, J. Guo, M. de Rijke, Y. Fan, X. Cheng, Are large language models good at utility judgments?, in: G. H. Yang, H. Wang, S. Han, C. Hauff, G. Zuccon, Y. Zhang (Eds.), Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, ACM, 2024, pp. 1941–1951. URL: https://doi.org/10.1145/3626772.3657784. doi:10.1145/3626772.3657784.

[32] Y. Xiao, Y. Jin, Y. Bai, Y. Wu, X. Yang, X. Luo, W. Yu, X. Zhao, Y. Liu, Q. Gu, H. Chen, W. Wang, W. Cheng, Large language models can be contextual privacy protection learners, in: Y. Al-Onaizan, M. Bansal, Y. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, Association for Computational Linguistics, 2024, pp. 14179–14201. URL: https://aclanthology.org/2024.emnlp-main.785.

[33] T. Diethe, O. Feyisetan, B. Balle, T. Drake, Preserving privacy in analyses of textual data (2020). URL: https://www.amazon.science/publications/preserving-privacy-in-analyses-of-textual-data.

[34] R. Xin, N. Mireshghallah, S. S. Li, M. Duan, H. Kim, Y. Choi, Y. Tsvetkov, S. Oh, P. W. Koh, A false sense of privacy: Evaluating textual data sanitization beyond surface-level privacy leakage, in: Neurips Safe Generative AI Workshop 2024, 2024. URL: https://openreview.net/pdf?id=3JLtuCozOU.

[35] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. M. Voorhees, Overview of the TREC 2019 deep learning track, CoRR abs/2003.07820 (2020). URL: https://arxiv.org/abs/2003.07820. arXiv:2003.07820.

[36] P. Ruch, C. Chichester, G. Cohen, F. Ehrler, P. Fabry, J. Marty, H. Müller, A. Geissbühler, Report on the TREC 2004 experiment: Genomics track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, volume 500-261 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2004. URL: http://trec.nist.gov/pubs/trec13/papers/uhosp-geneva.geo.pdf.

[37] F. L. De Faveri, G. Faggioli, N. Ferro, pypantera: A python package for natural language obfuscation enforcing privacy & anonymization, in: E. Serra, F. Spezzano (Eds.), Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024, ACM, 2024, pp. 5348–5353. URL: https://doi.org/10.1145/3627673.3679173. doi:10.1145/3627673.3679173.

[38] DeepSeek-AI, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,

2025. URL: https://arxiv.org/abs/2501.12948. arXiv:2501.12948.

[39] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, CoRR abs/2302.13971 (2023). URL: https://doi.org/10.48550/arXiv.2302.13971. doi:10.48550/ARXIV.2302.13971. arXiv:2302.13971.

[40] H. M. Culbertson, What is an attitude?, The Journal of Extension 6 (1968) 9.

[41] N. Schwarz, G. Bohner, The construction of attitudes, Blackwell handbook of social psychology: Intraindividual processes (2001) 436–457.

[42] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), 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, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410. doi:10.18653/V1/D19-1410.

[43] S. M. Jayanthi, V. Embar, K. Raghunathan, Evaluating pretrained transformer models for entity linking in task-oriented dialog, CoRR abs/2112.08327 (2021). URL: https://arxiv.org/abs/2112.08327. arXiv:2112.08327.