Inspecting the Representation Manifold of Differentially-Private Text

Stefan Arnold
Friedrich-Alexander-Universität Erlangen-Nürnberg
Lange Gasse 20, 90403 Nürnberg, Germany
stefan.st.arnold@fau.de

Abstract

Differential Privacy (DP) for text has recently taken the form of text paraphrasing using language models and temperature sampling to better balance privacy and utility. However, the geometric distortion of DP regarding the structure and complexity in the representation space remains unexplored. By estimating the intrinsic dimension of paraphrased text across varying privacy budgets, we find that word-level methods severely raise the representation manifold, while sentence-level methods produce paraphrases whose manifolds are topologically more consistent with human-written paraphrases. Among sentence-level methods, masked paraphrasing, compared to causal paraphrasing, demonstrates superior preservation of structural complexity, suggesting that autoregressive generation propagates distortions from unnatural word choices that cascade and inflate the representation space.

1 Introduction

Language Models (LMs) (Chowdhery et al., 2023) are trained on extensive corpora of text containing sensitive information. Several studies demonstrated that sensitive information can be extracted from LMs (Song and Shmatikov, 2019; Pan et al., 2020; Nasr et al., 2023; Carlini et al., 2023), raising significant privacy concerns and prompting the integration of privacy mechanisms.

To protect against unintended disclosure of information, *Differential Privacy* (DP) (Dwork et al., 2006) has been tailored to raw text (Fernandes et al., 2019; Feyisetan et al., 2020). Through a randomized mechanism, DP formalizes privacy through a notion of indistinguishability, ensuring that texts remain statistically unaffected by the addition or removal of individual samples in the text corpus.

While early randomized mechanisms exploit the distances between words in the embedding space (Mikolov et al., 2013) to replace words with a noisy approximation of their nearest neighbor, grammatical constraints associated with word-level privatization (Mattern et al., 2022) has led to a shift towards paraphrasing text at sentence-level by leveraging LMs (Igamberdiev and Habernal, 2023; Utpala et al., 2023; Meisenbacher et al., 2024).

Contribution. We inspect the representation geometry of text paraphrased under the privacy constraints of DP, accounting for different levels of privacy. Ansuini et al. (2019) discovered that high-dimensional signals reside on low-dimensional manifolds, a property that holds across neural representations (Tulchinskii et al., 2024). Building on *Intrinsic Dimensionality* (ID), we estimate the ID of texts and interpret ID shifts as a proxy for distortions on their structure and complexity. Specifically, we compare differentially-private transformations operating on word-level and sentence-level. We find that word-level DP deviates the most from human-authored paraphrases, significantly altering the underlying representation space. Concerning sentence-level DP, we argue that bidirectional paraphrasing based on masked substitution mitigates cascading errors that arise in sequential generation.

2 Background

We briefly provide the necessary foundations for differential privacy and intrinsic dimensionality.

2.1 Differential Privacy

Differential Privacy (DP) is a notion of privacy introduced by Dwork et al. (2006) under the term ε-indistinguishability. DP operates on the principle of adding noise calibrated to the sensitivity of adjacent datasets that differ by at most one record. The level of indistinguishability can be controlled by the privacy budget $\varepsilon \in (0, \infty]$, with declining privacy guarantees as $\varepsilon \to \infty$.

To mitigate the disclosure of authorship (Song and Shmatikov, 2019), DP is applied to perturb raw

Proceedings of the Sixth Workshop on Privacy in Natural Language Processing, pages 53–59

text either at word level or sentence level through noise injected into embedding models (Mikolov et al., 2013) and language models (Peters et al., 2018; Radford et al., 2018), respectively.

Word-level DP. Feyisetan et al. (2020) introduced a randomized mechanism in which a text is perturbed at the word level by mapping each word to another word located within a radius derived from an embedding space and governed by the privacy budget ε. This randomized mechanism was termed `MADLIB`. By scaling the notion of indistinguishability by a distance, `MADLIB` satisfies the axioms of metric DP (Chatzikokolakis et al., 2013). Despite many refinements regarding the preservation of utility (Carvalho et al., 2021; Xu et al., 2021b; Yue et al., 2021) and privacy (Xu et al., 2020, 2021a), `MADLIB` continues to suffers from syntactic errors (Mattern et al., 2022) and semantic drift (Arnold et al., 2023).

Sentence-level DP. Given the shortcomings of `MADLIB` and its recent refinements (Yue et al., 2021; Chen et al., 2023), researchers conceptualized the privatization of text as paraphrasing by utilizing sequence-to-sequence models (Bo et al., 2021; Krishna et al., 2021; Weggenmann et al., 2022; Igamberdiev and Habernal, 2023). Unlike word-level mechanisms, which perturb text on a word-by-word basis, sentence-level mechanisms paraphrase entire sentences. A defining characteristic shared is the injection of noise into the encoder representations, and learning of the decoder to generate fluent paraphrases while obfuscating stylistic identifiers that could otherwise compromise privacy.

Mattern et al. (2022) conjectured that temperature sampling in LMs can be interpreted as an instance of the exponential mechanism (McSherry and Talwar, 2007), where the scoring function corresponds to most probable word given a context. The probability of selecting a word follows the softmax distribution over the *logits*, which represent the likelihood of each word occurring in a given context. Since DP requires the sensitivity to be bounded, these logits are clipped in range.

Since paraphrasing is contingent upon the resemblance between the training text and the text subjected to privatization, Utpala et al. (2023) leverage the generalization capabilities of large-scale pretrained LMs to generate paraphrases via zero-shot prompting. Meisenbacher et al. (2024) depart from autoregressive generation and instead adopted the idea of temperature sampling to masked LMs. Unlike causal LMs, which sample text sequentially, this approach masks words and predicts its substitution bidirectionally from context.

2.2 Intrinsic Dimensionality

Grounded on the manifold hypothesis (Fefferman et al., 2016), the concept of intrinsic dimensionality characterizes the number of degrees of freedom for data in a representation space. Unlike extrinsic dimensionality, which corresponds to the overall dimensionality of the representation space, the intrinsic dimension (ID) corresponds to the minimum number of coordinates which are necessary to approximately capture the variability, revealing the structure and complexity of the manifold. This renders the ID as a geometric property (Valeriani et al., 2023) that describes how data points are distributed within the representation space.

Several methods have been developed to estimate intrinsic dimensionality, each differing in its underlying assumptions and formulations. Levina and Bickel (2004) uses maximum likelihood estimation to fit the likelihood on the distances from one point to each point within a *fixed* neighborhood structure. If the neighborhood is set too small in a dense region, the dimensionality might be underestimated. If the neighborhood is set too large in a sparse region, it might be overestimated. Farahmand et al. (2007) adapts the size of the neighborhood based on the geometry of the manifold.

Facco et al. (2017) exploits the expected ratio of distances between closest neighbors, observing that the distribution of distances of a point to its first neighbor is significantly smaller than to its second neighbor in lower dimensions, while in higher dimensions, the distance ratio is relatively close. By relying on the minimal information needed from the neighborhood, this approach alleviates the effects of variations in densities and curvatures within the manifold, providing stable ID estimates.

Recent studies have investigated how intrinsic dimensionality evolves and manifests through the layers (Ansuini et al., 2019), with connections to learning dynamics (Aghajanyan et al., 2021; Pope et al., 2021) and generalization (Birdal et al., 2021). Ansuini et al. (2019) demonstrated that data embedded in a high-dimensional space is progressively compressed into low-dimensional manifolds.

Table 1: Overview of prominent techniques for differentially-private text rewriting. Scope specifies whether the method applies DP at the word-level or sentence-level. Mechanism indicates the type of privacy mechanisms. Budget refers to the recommended range of the privacy budget. Approach describes the underlying substitution mechanism, including word embeddings, causal LMs, conditional LMs, or masked LMs. Fine-tuned specifies whether the LM was explicitly fine-tuned for paraphrasing or only leveraged pre-trained representations.

	Scope	Mechanism	Budget	Approach	Fine-tuned
Feyisetan et al. (2020)	Word-level	Exponential	~ 10	Word Embedding	no
Mattern et al. (2022)	Sentence-level	Exponential	~ 100	Causal LM	yes
Igamberdiev & Habernal (2023)	Sentence-level	Gaussian	~ 1000	Conditional LM	no
Utpala et al. (2023)	Sentence-level	Exponential	~ 100	Causal LM	no
Meisenbacher et al. (2024)	Sentence-level	Exponential	~ 100	Masked LM	no

3 Methodology

We aim to investigate how privacy-preserving transformations alter the geometry of paraphrases relative to those generated without privacy guarantees.

For our experiments, we utilize `MRPC` (Dolan and Brockett, 2005), a dataset containing sentence pairs labeled for semantic equivalence. We selected sentence pairs that provide a *reference* and *paraphrase* to ensure a controlled basis for assessing geometric distortions in representation subspaces.

3.1 Selection of Privacy Mechanisms

Table 1 outlines key characteristics of prominent approaches for differentially-private rewriting. To ensure comparability across privacy budgets, we focus on randomized mechanisms that implement the exponential mechanism. For word-level paraphrasing, we select `Madlib` (Feyisetan et al., 2020), which perturbs individual word in embedding space. For sentence-level paraphrasing, we select `DP-PARAPHRASE` (Mattern et al., 2022), `DP-PROMPT` (Utpala et al., 2023), and `DP-MLM` (Meisenbacher et al., 2024), covering causal and masked paraphrasing with temperate sampling. `DP-PARAPHRASE` and `DP-PROMPT` are powered by fine-tuned `GPT-2` (Radford et al., 2019) and pre-trained `LLaMA-3` (Touvron et al., 2023), respectively. `DP-MLM` employs `RoBERTa` (Liu et al., 2019). Table 2 presents an example sentence from `MRPC` along with its human-authored and differentially-private paraphrases.

3.2 Estimation of Intrinsic Dimension

Following Tulchinskii et al. (2024), we obtain embeddings for each word in a text using `BERT` (Devlin et al., 2019), treating each text as a point cloud of words spanning a manifold in the representation space. The ID of this point cloud is then estimated using `TwoNN` (Facco et al., 2017). To ensure that ID estimations reflect meaningful linguistic properties rather than artifacts of tokenization, we drop demarcation tokens as `<CLS>` and `<SEP>`. We also filtered short text sequences with less than 15 words and truncated long text sequences at 128 words. This stabilizes ID estimates by ensuring that estimations are based on sufficiently rich representations, while avoiding outlier effects from excessively short or long sentences.

Our investigation spans a range of privacy budgets $\varepsilon \in \{10, 15, 20, 25, 50, 100\}$, allowing us to weigh the geometric distortions with respect to the desired level of privacy. Since temperature sampling is probabilistic, we repeat the paraphrasing process three times per sample at each privacy level, ensuring robust ID estimations across multiple trials and reducing variance in the distortions.

4 Findings

Figure 1 presents the deviation in the number of ID as a function of the privacy budget. To establish a lower bound for ID shifts, we measure the ID difference between reference sentences and their human-authored paraphrases from `MRPC`. This yields an ID shift of approximately 0.12, indicating that naturally occurring paraphrasing introduces only minimal geometric distortions in the representation space. Any privacy-preserving transformation that deviates strongly from this baseline alters the structure and complexity of text representations beyond natural variation, potentially affecting readability.

Word-Level Perturbation. Since `MADLIB` is applied at word-level, its randomized mechanism perturbs words independently, disregarding sentence structure and grammatical coherence. This results in fragmented and disorganized text, a phe-

Table 2: Example from MRPC showing a sentence and its human-authored paraphrase. Note that differentially-private paraphrases at word-level are obtained using a privacy budget of $\varepsilon = 25$, whereas differentially-private paraphrases at sentence-level are obtained using a privacy budget of $\varepsilon = 100$.

Sentence	Amrozi accused his brother, whom he called " the witness ", of deliberately distorting his evidence.
Paraphrase	Referring to him as only " the witness ", Amrozi accused his brother of deliberately distorting his evidence.
Feyisetan et al. (2020)	Amrozi accused his brother , Tyler he warn the witness confined deliberately discolored muse evidence.
Mattern et al. (2022)	The person is Amrozi . aggression is evident even illustrates its extreme inflections over their close relative.
Utpala et al. (2023)	The witness had said his wife had left him when his wife was pregnant, his second daughter was not Alis.
Meisenbacher et al. (2024)	He alleged his nephew, whom he named _ the witness " of specifically distracting his testimony.

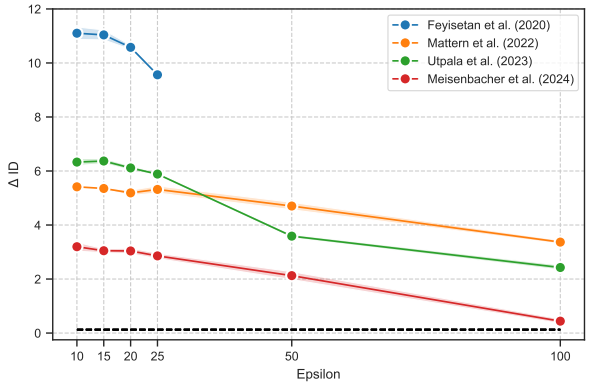

Figure 1: Shift in the estimated number of intrinsic dimensions, with a horizontal line representing a lower bound derived from human-authored paraphrases.

nomenon that can be observed through the highest ID shifts among all approaches. This observation reinforces a fundamental limitation of word-level perturbations, which induce severe distortions in representation subspaces, making them unsuitable for privacy-preserving paraphrasing.

Sentence-Level Perturbation. Unlike MADLIB, which perturbs words in isolation, sentence-level perturbation incorporates context when generating paraphrases. Across all privacy budgets, sentence-level perturbation introduces significantly less distortion, as indicated by their consistently lower ID shifts. This demonstrates that leveraging LMs produces more natural paraphrases.

Among causal paraphrasing, a mixed pattern emerges depending on the privacy regime. The ID shift of DP-PARAPHRASE remains stable across privacy budgets, whereas DP-PROMPT declines more sharply. At strict privacy regimes, DP-PARAPHRASE, which is explicitly fine-tuned for paraphrasing, outperforms DP-PROMPT, which learns paraphrasing implicitly from pre-training. At more relaxed privacy regimes, however, DP-PROMPT surpasses DP-PARAPHRASE by operating more within human-like representation geometry. Since privacy is enforced via temperature sampling, this trend suggests differing sensitivity to temperature values. DP-PARAPHRASE handles high temperatures more effectively, whereas DP-PROMPT tends to generate excessively complex paraphrases. Unlike autoregressive paraphrasing, DP-MLM adopts masked paraphrasing, reconstructing words bidirectionally rather than generating words sequentially. DP-MLM clearly excels across all privacy budgets, yielding more stable representation geometry.

Error Propagation We argue that a key factor driving the divergence between causal and masked paraphrasing stems from error propagation. Causal paraphrasing perturbs text in a fixed order, where each word conditions the selection of the next word, whereas masked paraphrasing operate bidirectionally, conditioning each word substitution on both preceding and following context. When differential privacy is enforced through temperature sampling, it introduces randomness, destabilizing generation by increasing the likelihood of unnatural word choices. Once a word has been poorly substituted, the language model must compensate to maintain fluency, leading to cascading errors which manifest in the form of drastic changes in the representation subspace. Since masked paraphrasing is not constrained by sequential consistency, distortion from a poorly chosen word does not propagate along the sentence, preventing error accumulation and producing more stable paraphrases.

5 Conclusion

We analyze the transformative effects of applying DP to text, focusing on how privacy constraints induce geometric distortions in the representation space. By leveraging the ID as a measure of structural complexity, we assess the extent to which prominent DP mechanisms alter latent subspaces and reshape linguistic representations. Our find-

ings reveal that word-level DP introduces severe ID shifts, leading to drastically inflated representation manifolds. For sentence-level DP, we observe distinct differences between their representation geometry, depending on how words are substituted and whether errors from suboptimal word choices accumulate and propagate throughout a sentence.

Limitations. A limitation of our inspection is that ID estimation, while a powerful tool for inspecting representation geometry of text, does not directly capture linguistic quality. Although ID shifts provide evidence of geometric distortions, connecting these distortions to measures of fluency (Salazar et al., 2020) and adequacy (Zhang et al., 2019; Yuan et al., 2021) would complement our understanding of alterations induced by DP rewriting.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. 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)*, pages 7319–7328.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023. Driving context into text-to-text privatization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 15–25, Toronto, Canada. Association for Computational Linguistics.

Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. 2021. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in Neural Information Processing Systems*, 34:6776–6789.

Haohan Bo, Steven H. H. Ding, Benjamin C. M. Fung, and Farkhund Iqbal. 2021. ER-AE: Differentially private text generation for authorship anonymization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3997–4007, Online. Association for Computational Linguistics.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ricardo Silva Carvalho, Theodore Vasiloudis, and Oluwaseyi Feyisetan. 2021. Tem: High utility metric differential privacy on text. *arXiv preprint arXiv:2107.07928*.

Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102. Springer.

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*, pages 5747–5758, Toronto, Canada. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140.

Amir Massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. 2007. Manifold-adaptive dimension estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 265–272.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer, Cham.

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*, pages 178–186.

Timour Igamberdiev and Ivan Habernal. 2023. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. ADePT: Auto-encoder based differentially private text transformation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2435–2439, Online. Association for Computational Linguistics.

Elizaveta Levina and Peter Bickel. 2004. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 867–881, Seattle, United States. Association for Computational Linguistics.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024. DP-MLM: Differentially private text rewriting using masked language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9314–9328, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. 2021. The intrinsic dimension of images and its impact on learning. *9th International Conference on Learning Representations, ICLR*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2024. Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, 36.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. 2023. Locally differentially private document generation using zero shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8442–8457, Singapore. Association for Computational Linguistics.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. 2023. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. 2022. Dp-vae: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 721–731, New York, NY, USA. Association for Computing Machinery.

Nan Xu, Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. 2021a. Density-aware differentially private textual perturbations using truncated gumbel noise. In *The International FLAIRS Conference Proceedings*, volume 34.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020. A differentially private text perturbation method using a regularized mahalanobis metric. *arXiv preprint arXiv:2010.11947*.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2021b. On a utilitarian approach to privacy preserving text generation. *arXiv preprint arXiv:2104.11838*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

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*, pages 3853–3866, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *International Conference on Learning Representations*.