arXiv:2506.09591v1 [cs.CL] 11 Jun 2025

Memorization in Language Models through the Lens of Intrinsic Dimension

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

Abstract

Language Models (LMs) are prone to memorizing parts of their data during training and unintentionally emitting them at generation time, raising concerns about privacy leakage and disclosure of intellectual property. While previous research has identified properties such as context length, parameter size, and duplication frequency, as key drivers of unintended memorization, little is known about how the latent structure modulates this rate of memorization. We investigate the role of *Intrinsic Dimension* (ID), a geometric proxy for the structural complexity of a sequence in latent space, in modulating memorization. Our findings suggest that ID acts as a suppressive signal for memorization: compared to low-ID sequences, high-ID sequences are less likely to be memorized, particularly in overparameterized models and under sparse exposure. These findings highlight the interaction between scale, exposure, and complexity in shaping memorization.

1 Introduction

Language Models (LMs) (Brown et al., 2020; Raffel et al., 2020; Chowdhery et al., 2023) are susceptible to memorizing segments of texts encountered during training (Shokri et al., 2017) and emitting these segments during generation (Nasr et al., 2025), even from corpora that has been subjected to deduplication (Kandpal et al., 2022; Lee et al., 2022). While memorization is connected to generalization (Arpit et al., 2017; Brown et al., 2021), it can cause severe issues such as inadvertent reproduction of personal information (Huang et al., 2022) and copyrighted materials (Lee et al., 2023).

To estimate memorization rates of LMs, Carlini et al. (2019) formalized a loose bound on memorization known as *exposure*, a metric that measures the relative difference in log-perplexity between *canaries*, synthetic sequences of text with fixed formats that are inserted during training and extracted during generation. By leveraging examples directly from the corpus, Carlini et al. (2023) introduced a tighter bound on memorization that avoids the need for canaries and reduces the computational overhead associated with computing exposure. Figure 1 visualizes the actionable methodology for examining memorization. Given a subset of examples, each split into a prefix p and a suffix s, memorization is estimated post-hoc by prompting the model f with the prefix and checking whether its continuation $f(p)$ replicates the reference s. The proportion of continuations that match the references verbatim provides an empirical estimate of memorization and quantifies the risk of information leakage.

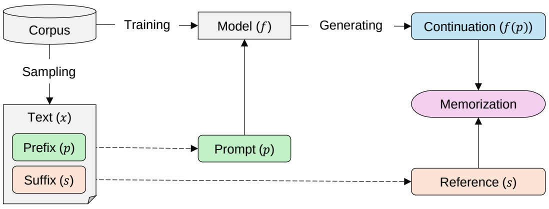


Figure 1: Overview of the post-hoc assessment of memorization, adapted from Kiyomaru et al. (2024). Methodologically, a sample x is split into a prefix p and a suffix s. By prompting p, the model f generates a continuation $f(p)$. If the continuation $f(p)$ matches s verbatim, the instance x is considered memorized.

Once memorization was evidenced in practice (Nasr et al., 2023), several properties have been identified as factors contributing to the memorization rate. Beyond its correlation with overfitting (Yeom et al., 2018), memorization is related to duplication counts (Carlini et al., 2023; Ippolito et al., 2023; Zhang et al., 2023; Kiyomaru et al., 2024), model capacity (Tirumala et al., 2022; Carlini et al., 2023), and context length (Carlini et al., 2023).

Grounded on the manifold hypothesis (Fefferman et al., 2016), few studies have examined the intrinsic dimension of data representations as a means to understand how neural networks structure latent spaces. These studies reveal that high-dimensional signals tend to lie in low-dimensional subspaces (Ansuini et al., 2019), and that intrinsic dimensionality acts as a geometric proxy for generalization capacity (Birdal et al., 2021; Pope et al., 2021).

Contribution. Assuming that the intrinsic dimension offers a lens onto sample complexity of sequences *as perceived* by language models, we investigate its relationship to the likelihood of memorization. Our investigation reveals that the intrinsic dimension systematically modulates memorization behavior: sequences with low intrinsic dimension, residing in compressed subspaces, are more amenable to memorization, particularly under sparse exposure, whereas sequences with high intrinsic dimension are less frequently memorized unless they are encountered repeatedly.

2 Background

We briefly provide necessary foundations for unintended memorization and intrinsic dimensionality.

2.1 Unintended Memorization

Memorization is commonly referred to the phenomenon of a neural network to fit arbitrarily assigned labels to features (Zhang et al., 2022). Although viewed as a sign of overfitting, memorization is linked to generalization (Arpit et al., 2017), particularly for data with long-tailed distributions (Feldman, 2020; Feldman and Zhang, 2020), where memorization can serve as an inductive bias that enables models to generalize beyond dominant modes and learn from rare or noisy examples.

Unintended Memorization, which refers to the reproduction of data used for training during generation, stands in contrast to these desirable forms of memorization (Brown et al., 2021). A longstanding belief held that memorization arises in the presence of overfitting (Yeom et al., 2018), however, this belief has been challenged by recent findings showing memorization in the absence of overfitting (Tirumala et al., 2022). Since large-scale language models have been found to memorize content even when trained on massively deduplicated text, overfitting only presents a sufficient condition but not a necessary condition for memorization.

Calling for a more nuanced understanding of unintended memorization, several notions have been operationalized. Depending on their degree of fidelity, these notions can be broadly categorized into *verbatim memorization*, in which sequences must match exactly, and *approximate memorization*, which allows for slight variations (Ippolito et al., 2023). Noteable definitions for memorization include *canary memorization* (Carlini et al., 2019), *eidetic memorization* (Carlini et al., 2021), *counterfactual memorization* (Feldman and Zhang, 2020; Zhang et al., 2023), *discoverable memorization* (Carlini et al., 2023; Hayes et al., 2024), and *distributional memorization* (Wang et al., 2025).

We adopt discoverable memorization as our actionable notion of memorization, formalizing the scenario in which a language model is prompted with the prefix of an example and is deemed to have memorized it if its continuation reproduces the suffix of the example *verbatim*. Carlini et al. (2023) operationalize this definition using deterministic decoding via greedy sampling, whereas Hayes et al. (2024) demonstrate its robustness across decoding strategies by accounting for temperature sampling.

2.2 Intrinsic Dimensionality

Unlike the ambient dimension of a representation space, the notion of *Intrinsic Dimension* (ID) characterizes the minimum number of latent directions required to represent data with minimal information loss (Fefferman et al., 2016). Geometrically, ID describes the manifold on which the data points are concentrated, capturing the effective dimensionality. The ID property has been used to gain insight into the sequential information flow in neural networks. Ansuini et al. (2019) showed that neural networks progressively compress high-dimensional data into low-dimensional manifolds, forming representations with orders-of-magnitude lower dimensionality than the ambient space.

A prototypical approach to estimate the ID involves projecting data onto a linear subspace (Jolliffe and Jolliffe, 1986). Since techniques relying on a linear projection poorly estimate the ID for data lying on curved manifolds, more recent techniques exploit local structures from nearest neighbors (Levina and Bickel, 2004; Farahmand et al., 2007; Facco et al., 2017; Amsaleg et al., 2018) or leverage the global topology (Schweinhart, 2021).

Levina and Bickel (2004) uses maximum likelihood estimation to fit a likelihood on the distances from a given point to its k-nearest points within a neighborhood structure. To stabilize ID estimations when confronted with variations in densities and curvatures within a manifold, Facco et al. (2017) considers only the ratio of distances between two closest neighbors, providing robust estimation from

Table 1: Examples of text and their corresponding number of dimensions occupied in latent space. Higher ID values indicate greater geometric complexity.

Text (truncated)	**ID**
We shall have no responsibility or liability for your visitation to, and the data collection and use practices of, such other sites. This Policy applies solely to the information collected in connection with your use of this Website and does not apply to any practices conducted offline or in connection with any other websites. [...]	2.08
Kazuni area there are many hippos and crocodiles which although rarely seen from the shore can certainly be heard at night. The location of the small town of Nkhata Bay is quite spectacular, a large, sheltered bay, accessible via a steep slope. Small boats transport the local people to various locations so that they can buy and sell, as there are hardly any roads around the lake. [...]	9.07

minimal neighborhood information. Schweinhart (2021) recently connects ID estimation to the well-established field of persistent homology by characterizing the continuous shape of the manifold at different scales to the upper box dimension. The upper box dimension is related to how efficiently points can be covered by boxes of decreasing size.

3 Methodology

We build on the setup introduced by Carlini et al. (2023) to assess memorization in relation to structural complexity. Specifically, we employ the GPT-neo model family (Wang and Komatsuzaki, 2021) and reuse their random sample derived from the Pile (Gao et al., 2020). To ensure that our measurements isolate structural complexity from confounding factors, we carefully control sequence length and duplication counts. We restrict all sequences to a uniform length of 150, thereby stabilizing ID estimations. We subsample $1,000$ sequences stratified by duplication frequency on a logarithmic scale for ranges between $[1, 10)$, $[10, 100)$, and $[100, 1000)$, allowing us to disentangle the influence of duplication from that of structural complexity.

To estimate the ID, we follow Tulchinskii et al. (2024) by treating each text as a point cloud spanning a manifold in the embedding space. We then obtain contextualized embeddings using BERT (Devlin et al., 2019), and estimate the intrinsic dimension using TwoNN (Facco et al., 2017), discarding artifacts of tokenization. Table 1 depicts example sequences and their corresponding IDs, which we interpret as a proxy for complexity in latent space.

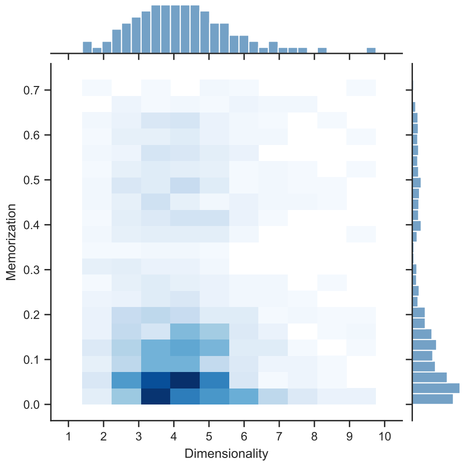


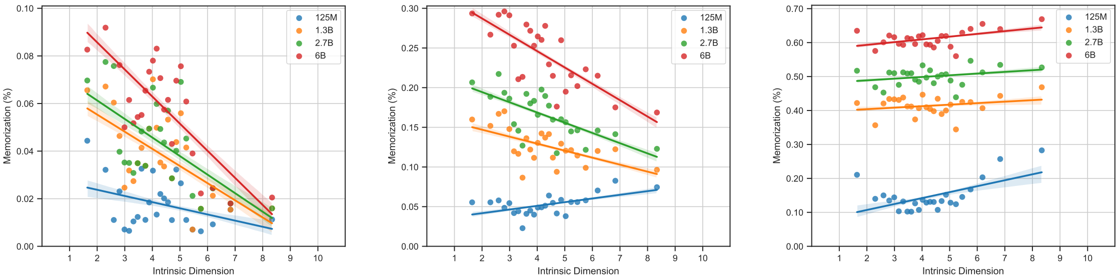
Figure 2: Distribution of memorization rate and intrinsic dimension, aggregated across scale and exposure.

Figure 2 shows the joint distribution of the memorization rate and intrinsic dimensionality, aggregated across model sizes and duplication counts. We observe that most samples cluster in regions characterized by low dimensionality and low memorization. However, when disaggregating by model scale and number of duplications, clear patterns emerge that elucidate the relationship between structural complexity and rate of memorization.

4 Findings

Figure 3 presents the relationship between memorization rate and intrinsic dimensionality for ascending levels of duplication frequency. Specifically, we quantile-binned the intrinsic dimension into 25 equally sized intervals and averaged memorization within each bin. Each subplot further disaggregates model capacity, covering models with roughly 0.1, 1.3, 2.7, and 6.0 billion parameters.

Consistent with the relationships reported by Carlini et al. (2023), our findings reveal a log-linear increase of memorization as a function of both duplication count and model capacity. Beyond these relationships, we observe a modulating influence of the intrinsic dimension. In the low-duplication regime, memorization declines inversely with intrinsic dimensionality across all model sizes. This inverse trend indicates that complex sequences, particularly those lying on more intricate manifolds, are less likely to be memorized under sparse exposure. In the medium-duplication regime, we notice diverging patterns depending on the model sizes. The inverse relationship largely persists for large models, albeit with a diminished effect. However,



(a) $[1, 10)$

(b) $[10, 100)$

(c) $[100, 1000)$

Figure 3: Memorization as a function of intrinsic memorization, binned into equally-sized intervals and disaggregated by model scale. 3(a) presents a low-duplication regime, comprising samples with duplications of at most 10. 3(b) presents a medium-duplication regime, comprising samples with duplication frequencies ranging from 10 to 100. 3(c) presents a high-duplication regime, comprising samples with duplications capped at 1000.

this is not the case for small models. Once duplications are sufficiently frequent for memorization, small models display a reversal in trend, exhibiting a slight increase in memorization with rising structural complexity. This divergence may reflect the limited capacity of certain models to generalize, leading to greater memorization of sequences that they fail to compress effectively. In the high-duplication regime, memorization undergoes a further shift as it saturates and becomes almost invariant to the intrinsic dimension. These findings suggest that under conditions of frequent exposure, memorization is increasingly governed by exposure and scale, overriding the modulating influence of structural complexity.

5 Conclusion

Building on the shared connection of memorization and intrinsic dimension to generalization, we introduce the intrinsic dimension as a complementary factor shaping the likelihood of memorization in language models. Specifically, we examine the relationship between memorization rate and the structural complexity of sequences in latent space, conditioned on model scale and exposure frequency. For sufficiently parameterized models and moderate levels of duplication, the intrinsic dimension act as a suppressive signal on memorization. A reversed trend can be seen for models with limited capacity, which tend to memorize structurally complex sequences even under moderate exposure.

Limitations. Despite controlling for duplication frequency, we focus exclusively on exact duplicates, omitting near-duplicates which are known to account for the majority of memorized content in large-scale corpora (Lee et al., 2022). This constraint likely underestimates memorization. Additionally, we restrict our analysis to verbatim memorization, a narrow definition that is known to give a false sense of privacy (Ippolito et al., 2023). Finally, we rely on greedy decoding to measure memorization, however, this decoding strategy is atypical in practical deployments (Hayes et al., 2024).

References

Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E Houle, Ken-ichi Kawarabayashi, and Michael Nett. 2018. Extreme-value-theoretic estimation of local intrinsic dimensionality. *Data Mining and Knowledge Discovery*, 32(6):1768–1805.

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.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR.

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.

Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. 2021. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*, pages 123–132.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

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.

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.

Vitaly Feldman. 2020. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, and Ilia Shumailov. 2024. Measuring memorization through probabilistic discoverable extraction. *arXiv preprint arXiv:2410.19482*.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. 2023. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53.

Ian T Jolliffe and IT Jolliffe. 1986. *Mathematical and statistical properties of sample principal components*. Springer.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.

Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. 2024. A comprehensive analysis of memorization in large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 584–596.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, pages 3637–3647.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

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

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*.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Florian Tramèr, and Katherine Lee. 2025. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning 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*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Benjamin Schweinhart. 2021. Persistent homology and the upper box dimension. *Discrete & Computational Geometry*, 65(2):331–364.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

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.

Ben Wang and Aran Komatsuzaki. 2021. Gpt-j-6b: A 6 billion parameter autoregressive language model.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2025. Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. In *The Thirteenth International Conference on Learning Representations*.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2022. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362.