On Data Contamination in Recommender Systems

Darío Garigliotti[1]

[1]University of Bergen, Norway

Abstract

The interest in studying the phenomenon of data contamination has recently increased due to the establishment of Large Language Models as the dominant technology for a vast array of information processing tasks. This **position paper** describes the reasons behind this increased awareness about data contamination and reflects on its possible implications for the field of Recommender Systems.

The phenomenon of *data contamination* happens in machine learning when one or more test instances have been incorporated to the set of training instances. This dataset that is used to train one or more machine-learnt models, and the eventual models themselves as well, are then considered *contaminated*, or *compromised* in the sense that the performance measurements obtained during model evaluation –and the conclusions built on top of them with respect to the hypotheses under consideration– do not necessarily correspond with the actual predictive power of the evaluated model [1]. Indeed, as test instances are part of the learnt distribution, model evaluation over those instances may exaggerate its claimed generalization capabilities [2]. Although the phenomenon was already known [3], the study of data contamination has significantly increased in recent times of explosion of generative AI with the dominance that Large Language Models (LLMs) have across several spaces of research problems in language technology and closely related interests from industrial applications [1, 4, 5]. The mechanisms for training LLMs are particularly affected by contaminated data in multiple manners. Also, although a possibly short-lived definition of data contamination might have referred to test instances *inadvertently* added to the training set, as recent studies show, there exist scenarios behind widely-used LLMs where contamination may be part of its standard refinement practices [2]. Moreover, recent literature also confirms that data contamination can no longer be attributed exclusively to exact "copies" of test instances in training set but also due to training instances that, while not part of test set, are sufficiently similar to some counterpart in this test set. This work brings focus on data contamination and reflects on its possible implications for areas such as Recommender Systems, whose methodologies increasingly rely on LLM technologies and hence are affected by the downstream contamination phenomenon [6].

In a fundamental way, LLMs are trained autoregressively over vast amounts of crawled web corpora. In the linguistic patterns that characterize all this data lies the core of the statistically learnt abilities that the LLMs (seem to) exhibit. Any of these textual patterns that has been memorized in the LLM likely helps the prediction performed by an LLM when it matches part of the input that elicits generation [2]. The LLM training is typically complemented with the integration of several datasets in a multi-task learning fashion. These datasets are usually selected as representative among the collections developed in research for the study of phenomena behind one or more well-established tasks. The multi-task integration of datasets into the training has been observed to contribute towards contaminating LLMs [7, 8]. The contamination effects extend to the scenarios where the training also incorporates documentation guidelines that indicate, by instructions and possibly also via examples, how to annotate instances in a data labeling experiment. Additionally, with the establishment of *LLM-as-a-service*, the prominent IT providers behind them have an incentive to identify instances that are convenient to be added into the training, typically via fine-tuning [2]. Examples of these instances are cases where the model underperforms, but also instances that the back-end algorithms of a closed, commercial LLM deems interesting, possibly due to sufficient rareness with respect to the data already incorporated.

These characteristics of the data contamination phenomenon make it a multi-faceted problem space that is currently under heavier scrutiny from the Natural Language Processing (NLP) community [1, 2, 9]. The literature reporting and quantifying aspects of (part of) the detected magnitude of contamination in an LLM is rather seldom [7, 10], which represents another key facet of the challenge for the involved research communities. Plus, the aforementioned technical discrepancies in the overclaims about generalizability that may arise by analyzing performances of contaminated models ramify into ethical consequences of various degrees. These implications can be particularly harmful in sensitive application areas

ROEGEN@RECSYS'24: The 1st Workshop on Risks, Opportunities, and Evaluation of Generative Models in Recommender Systems, Colocated with ACM Conference on Recommender Systems

dario.garigliotti@uib.no (D. Garigliotti)
0000-0002-0331-000X (D. Garigliotti)

such as the domains of medical, financial or legal decision making [11]. The same work also notes that the incentives for business organizations that provide digital services based on generative AI likely clash with possible awareness and admission about wrongly claimed abilities in these technologies. The possible ethical implications of the usage of contaminated models should be considered as extending the space of concerns under study in communities like Recommender Systems, where potential harms derived by generative methods have recently increased their presence in relevant literature [6].

In spite of the increasing efforts to refine the characterization of data contamination and the study of proposed strategies to detect and minimize it [12, 11], the principle of *presumable compromise* assumes that if it can be contaminated, data should be considered as it is already compromised [2]. This phenomenon indirectly spreads out to other technologies, since contamination is inherited by datasets or models built on top of LLMs [2] through the variety of configurations in training regimes across stages like pre-training and fine-tuning [13]. Hence, the study of contaminated data is of high relevance for fields like Recommender Systems, where the developed methods to address its research problems are increasingly more based on LLMs [14, 15, 6].

In the area of LLM-based recommender systems, at least two contamination sub-phenomena may require especial interest from the community. First is the aspect of partial overlap. Specifically, a test instance does not always need to fully appear in the train set, but instead it is enough for one of its components to have been present during training to possibly contribute to contamination [3]. As an example, if each instance in a dataset is composed of (certain representation for each component of) a user, an item and a score, the user component being already seen and memorized during training can help the LLM perform better in its predictive abilities, and by this it can then also lead to wrong claims about generalization abilities from these results. Secondly, the aspect of sufficient similarity allows for a training instance to be contributing to contamination if it is sufficiently similar to a test counterpart. In NLP studies, this has been observed in scenarios where an instance is similar to another in some linguistic space, e.g. one is a paraphrase or a translation of the other [16]. In the context of recommender systems, similarity could occur, paradigmatically, between respective instances which partially or fully characterize two users or two items by any of multiple possible representations (e.g. data based on interactions, user or item descriptions, queries, reviews, ratings, associated images, among others).

Another distinguished aspect in data contamination that should be taken into account by the Recommender Systems community is the employment of generative models to augment annotated data [17, 18, 19], especially the developments towards automatic data labeling with LLMs [20]. Whether the annotations are obtained by approaches like pseudo-labeling or weak supervision [21, 22] or by interaction-based data as in traditional recommender systems [6], these labeling strategies may introduce errors in several manners. And model families that are subject to be affected by contaminated data, particularly LLMs, can be detrimental in the ways associated to contamination that, as described here, are still in early study and still to be fully characterized.

This paper has described the recently increased interest on studying data contamination in generative AI as crucially dependable on LLMs, and its reflection on the space of possible implications for the Recommender Systems calls for the analysis about the need for a corresponding awareness from this field.

Acknowledgments

This work was funded by the Norwegian Research Council grant 329745 Machine Teaching for Explainable AI.

References

[1] O. Sainz, J. Campos, I. García-Ferrero, J. Etxaniz, O. L. de Lacalle, E. Agirre, NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 10776–10787. URL: https://aclanthology.org/2023.findings-emnlp.722. doi:10.18653/v1/2023.findings-emnlp.722.

[2] A. Jacovi, A. Caciularu, O. Goldman, Y. Goldberg, Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 5075–5084. URL: https://aclanthology.org/2023.emnlp-main.308. doi:10.18653/v1/2023.emnlp-main.308.

[3] P. Lewis, P. Stenetorp, S. Riedel, Question and answer test-train overlap in open-domain question answering datasets, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1000–1008. URL: https://aclanthology.org/2021.eacl-main.86. doi:10.18653/v1/2021.eacl-main.86.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[6] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasirzadeh, S. Milano, A review of modern recommender systems using generative models (gen-recsys), 2024. URL: https://arxiv.org/abs/2404.00579. arXiv:2404.00579.

[7] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, M. Gardner, Documenting large webtext corpora: A case study on the colossal clean crawled corpus, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1286–1305. URL: https://aclanthology.org/2021.emnlp-main.98. doi:10.18653/v1/2021.emnlp-main.98.

[8] Y. Elazar, A. Bhagia, I. H. Magnusson, A. Ravichander, D. Schwenk, A. Suhr, E. P. Walsh, D. Groeneveld, L. Soldaini, S. Singh, H. Hajishirzi, N. A. Smith, J. Dodge, What's in my big data?, in: The Twelfth International Conference on Learning Representations, 2024.

[9] B. Mehrbakhsh, D. Garigliotti, F. Martínez-Plumed, J. Hernandez-Orallo, Confounders in instance variation for the analysis of data contamination, in: O. Sainz, I. García Ferrero, E. Agirre, J. Ander Campos, A. Jacovi, Y. Elazar, Y. Goldberg (Eds.), Proceedings of the 1st Workshop on Data Contamination (CONDA), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 13–21. URL: https://aclanthology.org/2024.conda-1.2. doi:10.18653/v1/2024.conda-1.2.

[10] M. Riddell, A. Ni, A. Cohan, Quantifying contamination in evaluating code generation capabilities of language models, arXiv preprint arXiv:2403.04811 (2024).

[11] M. Ravaut, B. Ding, F. Jiao, H. Chen, X. Li, R. Zhao, C. Qin, C. Xiong, S. R. Joty, How much are llms contaminated? a comprehensive survey and the llmsanitize library, ArXiv abs/2404.00699 (2024). URL: https://api.semanticscholar.org/CorpusID:268819579.

[12] W. Zhu, H. Hao, Z. He, Y. Song, Y. Zhang, H. Hu, Y. Wei, R. Wang, H. Lu, Clean-eval: Clean evaluation on contaminated large language models, arXiv preprint arXiv:2311.09154 (2023).

[13] I. Magar, R. Schwartz, Data contamination: From memorization to exploitation, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 157–165. URL: https://aclanthology.org/2022.acl-short.18. doi:10.18653/v1/2022.acl-short.18.

[14] Z. Zhao, W. Fan, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang, Q. Li, Recommender systems in the era of large language models (llms), 2024. URL: https://arxiv.org/abs/2307.02046. arXiv:2307.02046.

[15] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, H. Xiong, E. Chen, A survey on large language models for recommendation, 2024. URL: https://arxiv.org/abs/2305.19860. arXiv:2305.19860.

[16] W. Zhu, H. Hao, Z. He, Y.-Z. Song, J. Yueyang, Y. Zhang, H. Hu, Y. Wei, R. Wang, H. Lu, CLEAN–EVAL: Clean evaluation on contaminated large language models, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 835–847. URL: https://aclanthology.org/2024.findings-naacl.53.

[17] Q. Liu, F. Yan, X. Zhao, Z. Du, H. Guo, R. Tang, F. Tian, Diffusion augmentation for sequential recommendation, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1576–1586. URL: https://doi.org/10.1145/3583780.3615134. doi:10.1145/3583780.3615134.

[18] Z. Wu, X. Wang, H. Chen, K. Li, Y. Han, L. Sun, W. Zhu, Diff4rec: Sequential recommendation with curriculum-scheduled diffusion augmentation, in: Proceedings of the 31st ACM International Conference on Multimedia, MM '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 9329–9335. URL: https://doi.org/10.1145/3581783.3612709. doi:10.1145/3581783.3612709.

[19] H. Zou, C. Caragea, JointMatch: A unified approach for diverse and collaborative pseudo-labeling to semi-supervised text classification, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 7290–7301. URL: https://aclanthology.org/2023.emnlp-main.451. doi:10.18653/v1/2023.emnlp-main.451.

[20] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, H. Liu,

Large language models for data annotation: A survey, 2024. URL: https://arxiv.org/abs/2402.13446. arXiv:2402.13446.

[21] M. Dehghani, H. Zamani, A. Severyn, J. Kamps, W. B. Croft, Neural ranking models with weak supervision, in: N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, R. W. White (Eds.), Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017, ACM, 2017, pp. 65–74. URL: https://doi.org/10.1145/3077136.3080832. doi:10.1145/3077136.3080832.

[22] H. Zamani, M. Dehghani, F. Diaz, H. Li, N. Craswell, Sigir 2018 workshop on learning from limited or noisy data for information retrieval, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1439–1440. URL: https://doi.org/10.1145/3209978.3210200. doi:10.1145/3209978.3210200.