On the implications of data contamination for Information Retrieval systems*

Darío Garigliotti[1][0000−0002−0331−000X]

University of Bergen, Norway - dario.garigliotti@uib.no

Abstract. Data contamination occurs when test instances have been compromised during a training stage of building a machine learning model. The consequences of this phenomenon over the quality of learning data are crucial when evaluating a learned predictor, since it could distort the assessment of the actual capabilities of the system. Its study has recently gained more traction in the research on Large Language Models, where it is common to chase performances in order to support claims about model abilities. Since the field of Information Retrieval increasingly studies and develops approaches that rely on these data-centric technologies, this position paper considers the phenomenon of data contamination in terms of its possible consequences for this field.

Keywords: Data Contamination · Information Retrieval · Position work

Data contamination is the phenomenon that occurs in machine learning when test instances have been made part of the set of instances used at training stage. Although already observed in the literature (Lewis et al., 2021), the study of data contamination by the Natural Language Processing (NLP) community has very recently risen significantly, due to the widespread success of Large Language Models (LLMs) as the dominant language technology in a variety of tasks (Sainz et al., 2023; Jacovi et al., 2023; Radford et al., 2019; Touvron et al., 2023). The long training stages of an LLM typically involve very large volumes of data mostly crawled from web pages, which provide vast amounts of text to use in autoregressive language modeling. This is complemented in multi-task learning fashion with the integration of prominent datasets that the research community has used to evaluate their models trained for multiple tasks. It is in these scenarios that test instances have been included in the training regimes of LLMs (Sainz et al., 2023). On the one hand, the learnt, often memorized, patterns over textual content in autoregressive modeling serve at the core of the text generation capabilities at inference time. Any test instance that matches part of the learnt content is possibly helped by the memorization in the model, and hence weakens any claim about generalization abilities in such a model, i.e. about the ability to predict for unseen cases (Jacovi et al., 2023). On the other hand, the incorporation of existing datasets in multi-task learning allows for inadvertently adding part of the test set into the training instances, as it has been observed in

* Part of NRF project 329745 Machine Teaching For XAI.

well-known LLMs (Dodge et al., 2021; Elazar et al., 2024). The problem does not limit to only the actual test instances being made part of training, but extends also to the situations where documentation or guidelines, of various levels of detail and possibly containing examples of labeled instances, about how to label data are incorporated (Sainz et al., 2023). Furthermore, in what a priori seems a less known phenomenon, an already-deployed, closed, commercial LLM like one of the GPT family has the incentives to incorporate —by some mechanism that involves fine-tuning— additional instances from users of its API that are deemed useful to improve the model (Jacovi et al., 2023). Examples of this kind of instances are those in which the model performs poorly —and whose processing should then be better handled— and/or those that are partially or fully identified as sufficiently dissimilar to the ones already used autoregressively —and would possibly allow for incorporating broader or more diverse content—. These natures of data contamination in language modeling —raw data during pre-training, training over datasets, and fine-tuning after deployment— (Sainz et al., 2023) cover a large set of potential scenarios where contaminated data is used. And moreover, data contamination encompasses a problem space for technologies in other related areas of research and application where NLP and machine learning models, prominently LLMs, are involved. Information Retrieval (IR) is one of these areas. This work presents a position about the importance of addressing the multiple aspects of data contamination in IR.

The Cranfield paradigm (Cleverdon, 1997) has characterized the long tradition in Information Retrieval of evaluating a system in a test collection. Such a collection compiles a set of information objects —usually documents— and a set of information needs —typically corresponding to a topic, expressed by a user query—, together with the judgments of relevance of each object for each need, all within a particular information seeking task (Voorhees, 2002). These test collections should provide the means for an evaluation in a controllable environment that allows for unbiased, reproducible experimentation on analyzing a retrieval system. The criteria considered to build an ideal test collection have historically required the need for a collection that has a large amount of documents and queries, from multiple sources multiply granular, with complete and sampled methods to obtain relevance judgments (Jones and van Rijsbergen, 1975). These criteria are behind strategies like pooling (Jones and van Rijsbergen, 1975; Harman, 2011), intended to collect large and diverse amounts of test instances, and by doing so, increase the chances of being representative and uncovering an unbiased sample, as well as of being able to reuse them (Büttcher et al., 2007). Assumptions and principles like these have marked the success of employing a test collection in such a strong evaluation tradition. They also overlap with the guiding forces and desires behind the construction of LLMs, namely, the incorporation of vast, diverse amounts of data with the expectation of obtaining a model that generalizes to as many tasks and as many unseen scenarios as possible. It could be argued that these common, well-oriented principles towards representativeness and generalization, across these related areas such as NLP and IR, might condition the involved technologies to be affected

by common phenomena, such as data contamination affects the achievement of those expectations (Sainz et al., 2023; Jacovi et al., 2023).

As machine learning approaches were incorporated into IR, the nature of a test collection lent itself to be used as a learning dataset by these approaches, often under strategies like k-fold cross validation given the rather small size of these collections to be split into sufficiently large training and testing subsets. Learning-to-Rank (LtR) serves as an underlying framework for many studied systems in IR, that has shown large success given its ability to combine the contributions of complementary methods in a learnt manner (Lucchese et al., 2019). The emergence of representation —or deep— learning and the corresponding Neural IR (NIR) paradigm (Mitra and Craswell, 2018) has seen a large body of literature studying its multiple aspects, within an umbrella of key assumptions, such as the following: (i) search based on distributional semantic signals reduces the traditionally problematic vocabulary gap, (ii) end-to-end differentiable learning of a method for a retrieval task avoids large part of the feature engineering, and (iii) vast amounts of data are needed for best exploiting NIR. With the recent developments in LLMs, yet another family of approaches comes to consideration to improve retrieval systems, from their implicit, parametric knowledge and from their ability to incorporate explicit information in frameworks like Retrieval-augmented Generation (RAG) (Lewis et al., 2020). These machine-learning-strong paradigms (LtR, NIR, LLM), and their increasing integration with each other, have significantly changed the nature of the retrieval systems, which are more data-driven, as well as have brought closer the practices and resources from other related research communities such as NLP, Computer Vision, and Reinforcement Learning. Several aspects that are encompassed by the phenomenon of data contamination, and that the last part of this work describes next, should bring more clarity on the extended space of implications within these integrated paradigms in Information Retrieval.

A distinguished theme is the possible incorporation of contaminated data during data annotation. The classic test collection paradigm previously discussed makes use of human annotators who provide judgments on the (degree of) relevance of a document for an information need, for each possible pair of considered documents and needs. Alternative approaches have been proposed to address the shortcomings of human labeling. Weakly supervision, for example, exploits signals considered weaker than those annotations, such as the scoring of an established ranking method like BM25, as label indicators of the relevance (Dehghani et al., 2017; Zamani et al., 2018). Another kind of signals correspond to those from user interaction, such as clicks, which can provide a mechanism to avoid the efforts in collecting explicit labels, as well as to overcome limitations in certain scenarios where privacy is sufficiently sensible (Zuccon, 2022). More recently, in correspondence with the widespread application of LLMs, there has been an increasing interest for using the implicit knowledge of LLMs as automatic annotators which could provide a truly vast number of useful, inexpensive judgments (Faggioli et al., 2023). The nature of the concept of relevance, that is minimal and independent (van Rijsbergen, 1979), contributes in its simplicity to

the possibility of judging such a relevance by means that result a priori convenient, yet could bring associated problems such as data contamination. Whether it involves editorial labels, interaction-based signals, or annotations via LLMs, these strategies are subject to make test instances part of the training stage. And especially with LLMs as labellers, the problem magnifies given the increasing awareness regarding the presence of contaminated data in those, and the incentives for the commercially driven ones to incorporate even more test data in their fine-tuning after deployment (Jacovi et al., 2023).

Additionally to these issues, a series of aspects within using contaminated data are important to be considered, such as the following items. In particular, a broader notion of contamination recently studied where test instances are *compromised* during training phase, rather than just being made part of, to capture observations about contamination beyond mere exact instance matching.

- The assumption of being *presumably compromised*: if it can be contaminated, it should be assumed as data already compromised (Jacovi et al., 2023).
- The property of *inheritance*: contamination of a dataset or model is inherited to every other model built on top of it (Jacovi et al., 2023).
- The observation that contamination can occur with partial instances in overlap (Lewis et al., 2021). For example, if each instance in a dataset comprises components such as a query, a document, a session, and a click signal, just one or more of these components memorized during training might contribute for a model to perform better in an overlapping instance at test time and not necessarily due to wrongly claimed generalization abilities of this model.
- The observation that a test instance does not need to match exactly one in training, but can instead being sufficiently similar such as the case of a paraphrase or a translation (Zhu et al., 2024).
- The ongoing research addressing detection, measurement, and avoidance of contaminated data (Sainz et al., 2023; Jacovi et al., 2023; Zhu et al., 2024).

These are relevant known dimensions of data contamination. This work scratches slightly beyond the surface of this problem in the space of research in an increasingly data-centric field like Information Retrieval, and calls for the community to be aware of the phenomenon and its possible implications in this field.

References

S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 63–70, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595935977. https://doi.org/10.1145/1277741.1277755. URL https://doi.org/10.1145/1277741.1277755.

C. Cleverdon. *The Cranfield tests on index language devices*, page 47–59. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. ISBN 1558604545.

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

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

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, and J. Dodge. What's in my big data? In *The Twelfth International Conference on Learning Representations*, 2024.

G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, and H. Wachsmuth. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23. ACM, Aug. 2023. https://doi.org/10.1145/3578337.3605136. URL http://dx.doi.org/10.1145/3578337.3605136.

D. Harman. *Information Retrieval Evaluation*. Morgan & Claypool Publishers, 1st edition, 2011. ISBN 1598299719.

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

K. S. Jones and C. J. van Rijsbergen. Report on the Need for and Provision of an "Ideal" Information Retrieval Test Collection. *British Library Research and Development Report 5266, Computer Laboratory*, 1975. URL https://cir.nii.ac.jp/crid/1570572699089480448.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

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

C. Lucchese, F. M. Nardini, R. K. Pasumarthi, S. Bruch, M. Bendersky, X. Wang, H. Oosterhuis, R. Jagerman, and M. de Rijke. Learning to rank in theory and practice: From gradient boosting to neural networks and unbiased learning. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1419–1420, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. https://doi.org/10.1145/3331184.3334824. URL https://doi.org/10.1145/3331184.3334824.

B. Mitra and N. Craswell. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126, 2018. ISSN 1554-0669. https://doi.org/10.1561/1500000061. URL http://dx.doi.org/10.1561/1500000061.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019. URL https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

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

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.

C. J. van Rijsbergen. *Information retrieval*. Butterworths, London, 2 edition, 1979. URL http://www.dcs.gla.ac.uk/Keith/Preface.html.

E. Voorhees. The philosophy of information retrieval evaluation. (2406), 2002-01-01 2002. URL https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=151546.

H. Zamani, M. Dehghani, F. Diaz, H. Li, and 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, page 1439–1440, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. https://doi.org/10.1145/3209978.3210200. URL https://doi.org/10.1145/3209978.3210200.

W. Zhu, H. Hao, Z. He, Y.-Z. Song, J. Yueyang, Y. Zhang, H. Hu, Y. Wei, R. Wang, and H. Lu. CLEAN–EVAL: Clean evaluation on contaminated large language models. In K. Duh, H. Gomez, and S. Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 835–847, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.findings-naacl.53`.

G. Zuccon. Pretrained language models rankers on private data: Is online and federated learning the solution? In O. Alonso, R. Baeza-Yates, T. H. King, and G. Silvello, editors, *Proceedings of the Third International Conference on Design of Experimental Search & Information REtrieval Systems, San Jose, CA, USA, August 30-31, 2022*, volume 3480 of *CEUR Workshop Proceedings*, pages 35–39. CEUR-WS.org, 2022. URL `https://ceur-ws.org/Vol-3480/paper-04.pdf`.