Department: AIML Column
Editor: Hsiao-Ying Lin, hsiaoying.lin@gmail.com

Unlearning in Large Language Models: We Are Not There Yet

Alberto Blanco-Justicia

Josep Domingo-Ferrer

Najeeb Moharram Jebreel

Benet Manzanares-Salor

David Sánchez
CYBERCAT Center for Cybersecurity Research of Catalonia, Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia

***Abstract*—The massive adoption of large language models (LLMs) has prompted concerns on how to align them with human ethics and the rule of law, and more precisely with data protection and copyright laws. Digital forgetting of undesirable knowledge via machine unlearning is a promising strategy whose progress and open problems we survey here.**

LARGE LANGUAGE MODELS (LLMs) have transformed the natural language processing (NLP) landscape by becoming the state of the art for most if not all tasks. However, since the release of ChatGPT, which brought the capabilities of LLMs to a broad audience, several issues have been raised, mainly regarding the alignment of LLMs with societal values and the rule of law. Such concerns include the impact of these models on the labor market, on the right to privacy of individuals, on copyright laws, on the furthering of biases and discrimination, and on the potential generation of harmful content.

Given the high cost and time required to train LLMs, retraining them from scratch to eliminate undesirable behaviors is often an impractical endeavor. Instead, there is a growing trend in the literature to adopt *machine unlearning* as an efficient means for digital forgetting in LLMs. Its objective is to transform a model with undesirable knowledge or behavior into a new model free of the detected issues without retraining from scratch. However, good unlearning mechanisms have to fulfill potentially conflicting requirements: (i) high unlearning effectiveness, that is, making sure the new model has forgotten the undesired knowledge/behavior; (ii) well-retained model performance on the desirable tasks; and (iii) reasonable cost and scalability.

Here we survey unlearning methods for LLMs by classifying them according to the scope and depth of the alterations performed on the target models. We also provide a summary of the main challenges faced in this area, and give recommen-

 Published by the IEEE Computer Society

dations for practitioners on how to choose specific methods depending on their needs.

UNLEARNING METHODS FOR LLMs

An essential requirement on any forgetting procedure is that it must allow checking the fulfillment of a forgetting request, particularly when such fulfillment is a legal obligation. We distinguish between *exact unlearning*, which guarantees that the undesired knowledge is no longer present in the updated model, and *approximate unlearning*, which only provides empirical evidences of forgetting.

We classify works in the literature into four primary categories: *global weight modification*, *local weight modification*, *architecture modification*, and *input/output modification*.

In *global weight modification*, every parameter of the model is subject to modification during the unlearning process. Methods in this category can employ *data sharding*, which consists of dividing the training data into multiple disjoint shards —each corresponding to a subset of the overall data— and training a separate model on each shard [1]. The outputs of the different shard models are then aggregated. Exact unlearning of a particular item can be achieved by suppressing the item from the shard containing it and retraining the model corresponding to that shard. Yet, these methods often entail substantial computational and time overheads, which renders them impractical for LLMs in many cases. Other mechanisms in this category include *gradient ascent* [2], which maximizes the loss on tokens to be forgotten, *knowledge distillation* [3], in which the unlearned model is treated as a student model that mimics a teacher model with desirable behavior, *reinforcement from human feedback* [4], which fine-tunes an LLM based on human feedback on what not to do, and *token fine-tuning* [5], which fine-tunes the model to erase original text.

In contrast, *local weight modification* affects only a specific subset of parameters. In this category we find methods employing *local retraining* [6], in which only LLM parameters relevant to target unlearning are optimized, methods that decompose LLMs into *task-specific vectors* that can be eliminated [7], and methods that locate and *directly modify* the parameters or neurons to unlearn a certain target [8]. These methods (and those mentioned above based on fine-tuning) are more efficient than those based on data sharding, but they are limited to approximate guarantees. Moreover, altering all or a substantial part of the target model's weights may degrade the model utility on the tasks to be retained.

To overcome this limitation, methods based on *architecture modification* add extra learnable parameters to the model [9] or use linear transformation [10] in order to achieve efficient and utility-preserving unlearning. Yet, if the architecture modifications can be reverted, the knowledge to be unlearned might be recovered again.

Finally, *input/output modification* methods only require access to the model inputs and outputs. This approach is especially useful when the operator has only black-box access to the model and thus cannot change its weights. In *input modification* [11], the forget input to an LLM is altered or adjusted to facilitate unlearning, whereas other methods selectively extract or manipulate the *information retrieved* from external sources to shape unlearning [12]; lastly, *in-context learning* exploits the in-context power of LLMs for unlearning [13]. These approaches can be deployed almost instantly to steer the model behavior as desired without significantly impacting the retain performance. However, they do not yield any real guarantee, as the models still retain the to-be-forgotten knowledge, which can be recovered by users using carefully crafted prompts. Although these approaches can serve as a fast hotfix to LLMs, they are only recommended as a short-term approach while waiting for other, more robust but more costly approaches to be applied.

CHALLENGES

A first challenge for practitioners is how to process forgetting or unlearning requests. This depends on the motivation for unlearning:

- Unlearning procedures aimed at removing private information are covered by many of the surveyed works. One reason is that privacy has been a research topic for several years, and definitions of privacy originally thought for databases have carried over into the machine learning field. Therefore, practitioners can choose from a wide range of methods,

which include those that provide exact guarantees but at high computing cost —*e.g.*, methods based on data sharding, whose set-up cost may be as high as retraining from scratch, even if they allow subsequent fast processing of unlearning requests—, as well as methods offering no guarantees but faster execution.

- Addressing copyright issues through unlearning is quite similar to the case of privacy. However, there is a difference between just ensuring that generative models do not output copyrighted material —to avoid plagiarism— and really forgetting/unlearning copyrighted material —needed when the rightholders forbid any processing of their protected work by the developers of an LLM. In the former case, the unlearning procedure only needs to filter the LLM's output to make sure it does not contain any verbatim fragments of a protected work. In the latter case, a fully fledged unlearning procedure such as those sketched above is needed.
- Unlearning requests can also be motivated by the poor quality of the LLM output: request of removal of factually incorrect information, fake news, or outdated information. This situation resembles that of privacy or copyright protection. Methods surveyed above that target *items* are applicable to address several unlearning motivations. The reason is that an item is a piece of data that can be as small as a token or as big as a whole document. Like for private or copyrighted data unlearning, the best choice of a method depends on the desired guarantees and affordable runtime.
- Ethical alignment is also a powerful motivation to request unlearning. Issues here include removal of biases and discrimination; stopping the generation of toxic, harmful, or hateful outputs; unlearning of other unwanted capabilities. Regarding mitigation of biases and discrimination, most of the literature just focuses on gender discrimination related to pronouns. Thus, more research is needed that covers additional aspects of discrimination and bias. Furthermore, most methods designed to correct biases offer little to no concrete guarantees, which forces practitioners to continuously monitor and assess any potential biases present in their models post unlearning.

A second challenge —already hinted in the previous paragraph— relates to forgetting guarantees. Methods giving such guarantees were mainly devised for general neural network classifiers. Their applicability to LLMs is not straightforward. On the one hand, guarantees refer to the parameters of the resulting models after training or forgetting. This means that, to obtain an *ex ante* guarantee of forgetting, a clear approximation of the influence of every training sample on the model weights must be known or computable. Whereas this may be feasible for small, simple models, it may be next to impossible for models as complex as LLMs. On the other hand, although unlearning an image or a record is relatively straightforward in classical machine learning, the unlearning complexity increases when dealing with unstructured text and LLMs. Unlearning in LLMs is harder than in traditional classification models due to the vast output space of language models, higher efficiency requirements, and limited access to the training data.

A third challenge relates to the evaluation of unlearning for LLMs. No *de facto* standard data sets exist to explicitly evaluate unlearning methods. Researchers often use specific data sets, which may fall short of yielding general enough evaluations. Further, all methods are evaluated on English-language data sets and there is no evidence of their performance for other languages. A greater number and a greater diversity (including language-diversity) of data sets on different unlearning applications are required to enable assessing how well unlearning methods generalize across different tasks and languages.

CONCLUSIONS

Unlearning is a promising approach to enforce digital forgetting in LLMs. However, if machine unlearning in general is a hot topic still far from maturity, this is even truer about machine unlearning in LLMs. The size and the unprecedented power of LLMs entail significant complexities is this still nascent area, but also exciting research opportunities.

ACKNOWLEDGMENTS

Partial support to this work has been received from Huawei Finland, the European Commission

(project H2020-871042 "SoBigData++"), the Government of Catalonia (ICREA Acadèmia Prizes to J.Domingo-Ferrer and to D. Sánchez, and grant 2021SGR-00115), MCIN/AEI/ 10.13039/501100011033 and "ERDF A way of making Europe" under grant PID2021-123637NB-I00 "CURLING" and European Union NextGenerationEU/PRTR via INCIBE (project "HERMES" and INCIBE-URV cybersecurity chair).

REFERENCES

1. L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 141–159.
2. J. Jang, D. Yoon, S. Yang, S. Cha, M. Lee, L. Logeswaran, and M. Seo, "Knowledge unlearning for mitigating privacy risks in language models," *arXiv preprint arXiv:2210.01504*, 2022.
3. L. Wang, T. Chen, W. Yuan, X. Zeng, K.-F. Wong, and H. Yin, "KGA: A general machine unlearning framework based on knowledge gap alignment," *arXiv preprint arXiv:2305.06535*, 2023.
4. X. Lu, S. Welleck, J. Hessel, L. Jiang, L. Qin, P. West, P. Ammanabrolu, and Y. Choi, "Quark: Controllable text generation with reinforced unlearning," *Advances in neural information processing systems*, vol. 35, pp. 27 591–27 609, 2022.
5. R. Eldan and M. Russinovich, "Who's Harry Potter? approximate unlearning in LLMs," *arXiv preprint arXiv:2310.02238*, 2023.
6. C. Yu, S. Jeoung, A. Kasi, P. Yu, and H. Ji, "Unlearning bias in language models by partitioning gradients," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 6032–6048.
7. G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," *arXiv preprint arXiv:2212.04089*, 2022.
8. X. Wu, J. Li, M. Xu, W. Dong, S. Wu, C. Bian, and D. Xiong, "DEPN: Detecting and editing privacy neurons in pretrained language models," *arXiv preprint arXiv:2310.20138*, 2023.
9. J. Chen and D. Yang, "Unlearn what you want to forget: Efficient unlearning for LLMs," *arXiv preprint arXiv:2310.20150*, 2023.
10. N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman, "LEACE: Perfect linear concept erasure in closed form," *arXiv preprint arXiv:2306.03819*, 2023.
11. I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, T. Yu, H. Deilamsalehy, R. Zhang, S. Kim, and F. Dernoncourt, "Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes," *arXiv preprint arXiv:2402.01981*, 2024.
12. E. Mitchell, C. Lin, A. Bosselut, C. D. Manning, and C. Finn, "Memory-based model editing at scale," in *International Conference on Machine Learning*. PMLR, 2022, pp. 15 817–15 831.
13. M. Pawelczyk, S. Neel, and H. Lakkaraju, "In-context unlearning: Language models as few shot unlearners," *arXiv preprint arXiv:2310.07579*, 2023.

Alberto Blanco-Justicia is an associate professor at Universitat Rovira i Virgili, Tarragona, Catalonia. He obtained a PhD in Computer Engineering from Universitat Rovira i Virgili. His interests are in data privacy, security, ethically-aligned design and machine learning explainability. Contact him at alberto.blanco@urv.cat.

Josep Domingo-Ferrer (Fellow, IEEE) received a PhD in computer science from the Autonomous University of Barcelona. He is a distinguished full professor of computer science and an ICREA-Academia researcher with Universitat Rovira i Virgili, Tarragona, Catalonia, where he also leads CYBERCAT. His research interests include data privacy, data security, and ethics-by-design. Contact him at josep.domingo@urv.cat.

Najeeb Moharram Jebreel is a senior posdoc researcher in Computer Science at Universitat Rovira i Virgili, Tarragona, Catalonia. He obtained a PhD in AI and Computer Security from Universitat Rovira i Virgili. His interests are in reconciling accuracy, security and privacy in distributed machine learning. Contact him at najeeb.jebreel@urv.cat.

Benet Manzanares-Salor is a PhD candidate at Universitat Rovira i Virgili, Tarragona, Catalonia. His research interests are in text anonymization and language models. Contact him at benet.manzanares@urv.cat.

David Sánchez (Senior Member, IEEE) is a full professor and an ICREA-Academia researcher with Universitat Rovira i Virgili, Tarragona, Catalonia. He obtained a PhD in computer science from the Technical University of Catalonia, Barcelona. His research interests include data semantics, machine learning, and

data privacy. Contact him at david.sanchez@urv.cat.