Thesis Proposal: Targeted and Unified Cross-Lingual Unlearning from Multilingual Language Models

Jan Bronec
Charles University
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague
Czech Republic
bronec@ufal.mff.cuni.cz

Jindřich Helcl
University of Oslo
Language Technology Group
Gaustadalleén 23 B, N - 0373 Oslo
Norway
jindrich@ifi.uio.no

Abstract

As large language models (LLM) trained on massive corpora scraped from the web exhibit the capability to reproduce sensitive and copyright-protected data, the field of machine unlearning has emerged to address the arising ethical and legal concerns. While previous research has provided a unified evaluation of LLM unlearning methods, this unification remains constrained to English-only models and datasets. We aim to address the prevailing fragmentation in recent cross-lingual unlearning research by extending existing unified benchmarks with multilingual data. To that end, we plan to compile a dataset of parallel translations of question-answer pairs consisting of real-world facts and synthetic personally identifiable information. Moreover, we will focus on mitigating model degradation during unlearning by selectively editing only those layers that contain the given knowledge.

1 Introduction

Transformer-based large language models (LLM; Vaswani et al., 2017) have been widely adopted in many natural language processing tasks thanks to their deep understanding of semantic structure. To acquire such understanding, training LLMs requires a massive amount of natural language data, which is often scraped from internet forums and other openly accessible sources. The sheer volume of training data rules out manual curation, and so maintainers of open training datasets, such as HPLT (Burchell et al., 2025) or Dolma (Soldaini et al., 2024), often rely on heuristic approaches to filter out unwanted content. Moreover, many proprietary and open-weight models have been trained on data that has not been publicly released (Touvron et al., 2023; Grattafiori et al., 2024), making it nearly impossible to publicly scrutinize the models' exposure to copyrighted, harmful, or otherwise unwanted content. This has raised both ethical and legal issues, as LLMs are prone to memorizing and recalling portions of their training data (Lee et al., 2022; Eldan and Russinovich, 2023).

Many production and open-weight models have been shown to regurgitate word-for-word not only copyright-protected creative content (Cooper et al., 2025; Ahmed et al., 2026), but also sensitive personally identifiable information (PII) (Carlini et al., 2021), which includes social security numbers, phone numbers, and people's locations. The former breaks the rules of *fair use*, as word-for-word copying of protected works is not considered transformative. This has led to numerous lawsuits against generative AI companies for training their models on datasets such as the Pile (Gao et al., 2020), which contains copyright-protected works.

The latter not only violates Differential Privacy (Dwork, 2006), but also conflicts with regulations such as the EU's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). As per the *"right to be forgotten"*, maintainers of these models may be obliged to remove this PII from their models, since it is often learned from web-scraped content without the consent of the individuals it corresponds to.

While removing this information from the training sets when it is found and retraining the models from scratch on the sanitized dataset might seem like a straightforward solution, retraining these large models on each data removal request is cost-prohibitive. The field of machine unlearning (Nguyen et al., 2024) aims to address the removal of such information by modifying models post-training. The goal of machine unlearning is to remove the model's knowledge of selected samples without significantly affecting its utility or fluency. Further problem definitions address various real-world scenarios. For example, continual unlearning (Gao et al., 2025) expects unlearning requests to emerge continuously and must be addressed each time.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 554–562

Although knowledge unlearning from LLMs has recently attracted considerable research, it faces several unresolved issues. Most importantly, despite many attempts, all state-of-the-art unlearning methods that sufficiently remove undesired knowledge also significantly degrade the models' fluency and utility. (Dorna et al., 2025; Ramakrishna et al., 2025b). Secondly, despite attempts to unify unlearning research (Dorna et al., 2025), we still observe strong fragmentation in the cross-lingual unlearning field. As different works utilize different datasets and metrics, comparing them remains difficult.

Thesis objectives. Our main objective is to improve upon existing unlearning methods and to mitigate model degradation under many unlearning requests. We ask the following question: "Can we improve the stability of LLM unlearning methods by localizing different kinds of knowledge and modifying only selected parts of the model?"

Furthermore, while there have been recent leaps in standardizing the benchmarking of unlearning methods and metrics (Dorna et al., 2025), the focus has been on English-only datasets and models. We aspire to further expand this standardization to cross-lingual unlearning.

Thesis structure. We dedicate the first part of the thesis, described in Section 2, to exploring targeted knowledge removal techniques. We will use interpretability techniques to identify model layers that are critical for conveying different kinds of knowledge. With a better understanding of where each type of information is stored, we will target specific layers for knowledge removal. Our goal is to mitigate the loss of model utility and increase the unlearning robustness to paraphrased prompts.

In the second part of the thesis, described in Section 3, we will focus on unifying the multilingual unlearning research, evaluating current state-of-the-art methods on unified benchmarks. To reinforce our comparisons and facilitate further research, we plan to compile a new dataset for cross-lingual unlearning, extending the research to a broader set of languages. We will generate this dataset of parallel translations automatically from public sources and by synthesising fictitious PII. Some of our target languages may require machine translation tools. In that case, we will ensure the quality of these automatic translations by having human annotators review them.

Applications. Complete retraining of LLMs on cleaned data will likely always be necessary to preserve the models' utility and ensure provable removal of unwanted knowledge. Despite that, machine unlearning has the potential to alleviate a substantial portion of the model maintenance cost by lowering the required frequency of retraining runs. A sufficiently stable unlearning method would allow granular knowledge removal updates to an existing model before the accumulation of removal requests warrants a full retraining. Moreover, the presence of sensitive information within pretrained models concerns not only the maintainers of systems such as chatbots or machine translation engines. A lot of research is done using openly accessible, pre-trained, open-weight models from repositories such as Hugging Face. However, many of these models were not trained on open datasets, as is the case, for example, with the Llama family of models (Touvron et al., 2023; Grattafiori et al., 2024). Because the training sets of these models cannot be publicly scrutinized for the presence of sensitive information, applications that use them may be susceptible to leaking it. In such cases, it is key to evaluate and reduce the presence of such data within the models.

2 Targeted Unlearning

2.1 SoTA Unlearning Methods

LLM unlearning has recently seen a surge of research, with many proposed unlearning methods. The majority of those methods are based on Gradient Ascent (GA; Jang et al., 2023), which involves standard fine-tuning of the entire model against tailored loss functions, aiming to suppress undesirable outputs. Examples of these are Negative Preference Optimization (NPO; Zhang et al., 2024), Simplified Negative Preference Optimization (SimNPO; Fan et al., 2025), or Unlearning via Self-Distillation with Adjusted Logits (UNDIAL; Dong et al., 2025). Instead of adapting the entire model, methods, such as Representation Misdirection for Unlearning (RMU; Li et al., 2024), attempt to disrupt the hidden representations of selected layers, which should correspond to undesired associations.

Despite numerous attempts to mitigate model degradation (Wang et al., 2025; Zhang et al., 2024; Fan et al., 2025), every currently available method that unlearns a measurable portion of requests also significantly degrades the model's utility (Ramakrishna et al., 2025b; Dorna et al., 2025). The degra-

dation issue is further highlighted in the continual unlearning scenario (Gao et al., 2025), where we expect unlearning requests arriving over time and need to continually remove them.

2.2 Stability Through Targeted Removal

Many of the aforementioned state-of-the-art LLM unlearning methods treat the model as a black-box, disregarding its underlying architecture and actual location of the specific knowledge. However, reasoning explainability techniques, such as mechanistic interpretability (Elhage et al., 2021) and causal mediation analysis (Meng et al., 2022), shed light on the models' inner workings. Mechanistic interpretability attempts to reverse-engineer the behavior of a transformer model by identifying "circuits" that correspond to actions it performs, such as copying a token from the input. Causal mediation analysis identifies critical layers within a model that mediate the recall of factual associations.

We take inspiration from the locate-then-edit approaches used by knowledge editing methods, such as Rank-One Model Editing (ROME; Meng et al., 2022) and Mass-Editing Memory in a Transformer (MEMIT; Meng et al., 2023). Meng et al. (2022) utilize causal mediation analysis to argue that certain factual associations can be attributed to a selected subset of feed-forward matrices within the larger model. We hypothesize that targeting these layers for knowledge removal updates should leave the model's overall behavior unaffected and help mitigate the deterioration of its utility. The RMU unlearning technique explores a similar idea by finetuning selected layers at specific tokens to produce noisy embeddings, thus disrupting unwanted associations.

Another feasible approach could involve applying knowledge editing methods to knowledge unlearning directly. Implementing this, however, is not trivial, as knowledge editing relies on specially annotated datasets. Methods like MEMIT compute the parameter edit for a chosen token within a longer context, corresponding to a subject. The context generally represents a relationship between the subject and another object. For example: "Michael Jordan plays basketball." The goal is then to change the model's association of the subject with a new object, like "football." Unlearning datasets, on the other hand, generally consist of unstructured input-output pairs without subject/object annotations (Maini et al., 2024; Shi et al., 2024; Li et al., 2024; Ramakrishna et al., 2025a).

Recent knowledge editing methods, such as AnyEdit (Jiang et al., 2025) and UnKE (Deng et al., 2025), focus on editing long-form, unstructured knowledge and have achieved promising results. Nonetheless, these methods still require replacement phrases for a given prompt. Li et al. (2025) circumvent this by providing a variety of "I don't know" phrases as the replacement output for ROME, MEMIT, GRACE (Hartvigsen et al., 2023), WISE (Wang et al., 2024), and AlphaEdit (Fang et al., 2025) editing methods.

It is a matter of discussion, as well as a concrete problem definition, whether forcing the model to respond with "I don't know" phrases to prompts about a specific person's PII solves the issue. In order to refuse an answer, knowledge of that person still has to be present in the model. Besides technically not resolving the issue of differential privacy, such overfit responses may open the door to membership inference attacks (MIA; Duan et al., 2024). MIAs are especially a problem for open-weight models, for which it is possible to infer which logits have been suppressed.

There are also alternative approaches, such as external model output content filtering safeguards, used for example in chat services by OpenAI (Markov et al., 2023) and DeepSeek (Guo et al., 2025). Besides being vulnerable to a variety of jailbreak attacks (Zhang et al., 2025a; Chu et al., 2025), these approaches also cannot offer protection against knowledge extraction from open-weight models.

2.3 Proposed Approach

A commonly used problem definition of knowledge unlearning that also satisfies differential privacy goes as follows. Assume a pre-training dataset $\mathcal{D} = \mathcal{D}_r \sqcup \mathcal{D}_f$ consisting of general knowledge $\mathcal{D}_r$ and unwanted knowledge $\mathcal{D}_f$. We are given a model $\pi_\theta(y|x)$ which was pre-trained on the entire dataset $\mathcal{D}$, and our goal is to find parameters θ', such that $\pi_{\theta'}(y|x)$ is consistent with a model trained purely on the dataset $\mathcal{D}_r$. Finetuning-based unlearning methods tend to be better aligned with this goal, as their loss functions incentivize the model to suppress unwanted outputs, increasing the likelihood of generic, uninformed answers. We propose combining current state-of-the-art knowledge editing methods with state-of-the-art finetuning-based knowledge unlearning methods.

Consider some context sequence x of n tokens $x_1, x_2, \ldots, x_n$, for which a given model π_θ gener-

ates an output sequence y. In general, knowledge editing methods treat a layer, or a selection of layers, as an *associative memory*, where for an input *key* vector k, the associative memory produces an output *value* vector v. The key vector k_i for each token x_i corresponds to the hidden state computed from the layers preceding the associative memory. Likewise, the value vector v_i is the hidden state computed by layers within the associative memory.

The goal of knowledge unlearning methods is to update the parameters of this associative memory in such a way that the updated model $\pi_{\theta'}$ generates a preferred output sequence y' corresponding to a new factual association.

Both structured and unstructured knowledge editing methods generally choose a single context token x_i for the associative memory update. For structured knowledge editing methods, the chosen token x_i typically corresponds to the subject within the context sentence. Unstructured knowledge editing methods often use the last token of the context.

Instead of fine-tuning the associative memory parameters, methods like ROME, MEMIT, and AlphaEdit first compute the new value vector v_i^* that the memory should produce for the key vector of token x_i in order for the model to produce the preferred output y'. The new value vector v_i^* is thus chosen through gradient descent such that it minimizes the negative log-likelihood (NLL) of the new fact y' being generated:

$$v_i^* = \arg\min_{v_i} -\log \pi_\theta(y'|x, h_i^l := v_i) \quad (1)$$

where the hidden state h_i^l for token x_i computed by the last layer l of the associative memory is being set to v_i before proceeding to the next layers. The methods then use v_i^* to directly apply an update to the associative memory, but they differ in how the update is performed.

Rather than optimizing the NLL loss, we can look for a value vector v_i^* that best suppresses the unwanted output y, eliminating the need for a replacement phrase. For example, we should be able to find v_i^* by optimizing for the NPO unlearning loss, assuming sufficient regularizations are applied as well:

$$v_i^* = \arg\min_{v_i} \mathcal{L}_{\text{NPO}}(v_i) \quad (2)$$

$$\mathcal{L}_{\text{NPO}}(v_i) = \frac{2}{\beta} \log \left(1 + \left(\frac{\pi_\theta(y'|x, h_i^l := v_i)}{\pi_\theta(y'|x)}\right)^\beta\right) \quad (3)$$

where β is a regularization hyper-parameter.

We expect that while unstructured editing methods like UnKE or AnyEdit will be useful for unlearning unstructured knowledge, we will need to take further measures to mitigate the utility degradation incurred by the edits. Experimenting with different unlearning losses, improving editing performance across different kinds of knowledge by targeting different parts of the model, and exploring similar approaches will constitute a substantial part of our work.

Li et al. (2026) explore a similar approach by targeting a single language-agnostic layer identified through Centered Kernel Alignment (Kornblith et al., 2019), and Linguistic Regions Development Score (Zeng et al., 2025). On this layer, they apply finetuning-based unlearning using RMU, SimNPO, and SLUG (Cai et al., 2024).

Nevertheless, fine-tuning for refusal responses may still be appropriate, for example, to mitigate the spread of hazardous knowledge, such as the creation of biological weapons or illegal drugs (Li et al., 2024). This knowledge can be inferred by the models, even when it is not directly present in the data, through the extrapolation of general and expert knowledge in fields such as biology and chemistry. Suppressing general knowledge of these fields would undermine the model's utility for legitimate uses as well. Rather, in this case, the model should be aware when to refuse an answer.

3 Multilingual Evaluation

Evaluating unlearning is a nuanced task involving knowledge, privacy risks, and utility desiderata. As such, it poses a major challenge in the field. Several benchmarks, such as TOFU (Maini et al., 2024), MUSE (Shi et al., 2024), WMDP (Li et al., 2024), and LUME (Ramakrishna et al., 2025a), have been proposed. Each comes with its own datasets, ranging from real and synthetic biographies containing PII, long-form creative documents, to multiple-choice questions involving hazardous knowledge.

However, each also comes with a different set of evaluation metrics, with no consensus on the best evaluation setup. This has fragmented the LLM unlearning research and made it difficult to compare the efficacy of different unlearning methods. Furthermore, many existing evaluation metrics have amassed considerable criticism, questioning their usefulness (Kim et al., 2025; Lynch et al., 2024;

Zhang et al., 2025b).

More recently, Dorna et al. (2025) used their OpenUnlearning framework to conduct a meta-analysis of commonly used unlearning evaluation metrics. They identified a set of metrics with reasonable efficacy and robustness. While those metrics still show room for improvement, we will be able to utilize their framework to ensure a sound comparison of our methods against competing works. We will also use this framework to test the soundness of future metrics.

While OpenUnlearning is a major step towards standardized unlearning evaluation, its work has so far been limited to English-only benchmarks. We are particularly interested in cross-lingual knowledge unlearning, an area where current research remains highly fragmented, especially due to the lack of standardized datasets.

Choi et al. (2024), for example, made use of the FLORES-200 dataset (Team et al., 2022) consisting of parallel translations of web articles, and BMLAMA53 (Qi et al., 2023) consisting of factual knowledge questions with single-word answers translated to different languages. Using these datasets, they evaluate cross-lingual unlearning with a regularized version of GradDiff (Liu et al., 2022) and negative task vector editing (Ilharco et al., 2023). Li et al. (2026) use MMMLU (OpenAI), which is a multilingual extension of the MMLU dataset (Hendrycks et al., 2021), focusing on multiple-choice factual questions across diverse fields. None of these datasets, however, focuses on PII data. More recently, Farashah et al. (2026) used the Google Translate API to translate the unlearning dataset TOFU (Maini et al., 2024) into nine other languages, and used it together with the SeeGULL dataset (Bhutani et al., 2024) to evaluate GradDiff and NPO.

Nevertheless, we plan to unite recent research, evaluating a wider range of unlearning methods on the same multilingual datasets for a clear comparison. Furthermore, we plan to build a new multilingual knowledge-unlearning dataset, specifically curated for cross-lingual unlearning, combining real-world knowledge with synthetic data and spanning a wider range of European languages.

Initially, we will focus on a limited set of European languages, for which we can ensure high-quality annotation through our academic peers. This set will include, but will not be limited to, Czech, Slovak, Norwegian, and Greek languages. This initial set will help us analyze the effect of language similarity on cross-lingual unlearning by including a mutually understandable language pair as well as a language written in a separate script. Following this initial analysis, we will aim to further extend our datasets to a wider set of diverse languages and scripts. For the initial analysis, we will be able to acquire annotations for a dataset of 1000 samples per language while ensuring fair compensation for our annotators.

To facilitate this, we will use open sources, such as Wikimedia data, to create a real-world, factual knowledge dataset of parallel translations and paraphrases. This part of the dataset shall be used for evaluating pre-training knowledge and its removal. For the second part of the dataset, we will focus on generating synthetic PII data, consisting of input-output pairs in different forms translated into a range of languages. Alternatively, we can also continue the work of Farashah et al. (2026) by extending their set of TOFU translations. For both parts, we will conduct quality evaluations through human annotation to ensure accurate translation.

Since the latter part will require fine-tuning models first, before attempting their unlearning, we will build a cross-lingual unlearning benchmark, focusing on state-of-the-art multi-lingual language models, such as Apertus (Apertus et al., 2025), EuroLLM (Martins et al., 2025), OpenEuroLLM,[1] and Aya Expanse (Dang et al., 2024). We plan to release all fine-tuned models for reproducibility and further research, and integrate our work into the OpenUnlearning framework.

4 Conclusion

In this research proposal, we have discussed the unsatisfactory reliability and robustness of current knowledge unlearning methods for large language models. As these methods degrade the fluency and utility of these models, we focus on developing more robust methods. We draw on the knowledge editing paradigm and propose targeting selected parts of the model, identified using interpretability techniques, as critical for conveying the given knowledge. By augmenting unlearning techniques with targeted editing, we hope to increase the utility of unlearned models and their robustness against paraphrased prompts and those in different languages.

Since the current cross-lingual research is difficult to compare because each uses a different set

[1] https://openeurollm.eu/

of datasets and evaluations, we plan to unify current research by evaluating unlearning methods in a unified cross-lingual benchmark on state-of-the-art multilingual language models. For that purpose, we propose compiling a new unlearning dataset composed of human-validated parallel translations of both real-world facts and synthetic personally identifiable information.

We hope our work will improve the comparison of existing unlearning methods in the cross-lingual setting and open the door to further research.

Limitations

Our upcoming work is subject to risks imposed by the complexity of the task. First, since we aim to utilize a targeted approach for the removal of different kinds of factual knowledge from the model, we rely on the efficacy of interpretability techniques such as causal mediation analysis to localize layers representing the given knowledge. Previous works have shown such methods to be reasonably effective at localizing layers critical to conveying short-form factual associations. However, these methods may not generalize across model architectures and different kinds of learned information, such as implicit facts and long-form memorized sequences.

Furthermore, our multi-lingual extension to the unlearning benchmarks will initially focus on a limited set of mostly mid-resource European languages. This focus may limit the generalizability of our findings to low-resource, non-Indo-European, and non-Latin-script languages.

Lastly, the quality of the synthetic component of our dataset depends on the realism of the generated data and the rigorousness of the annotators. Synthetically constructed entities, associations, and facts may not fully capture the diversity and context of real-world PII, possibly leading to misleadingly optimistic unlearning evaluations. As the field has not yet converged on universally accepted metrics and indicators, our evaluations remain subject to the limitations of existing benchmarks.

Acknowledgments

This work was supported by the project "Human-centred AI for a Sustainable and Adaptive Society" (reg. no.: CZ.02.01.01/00/23_025/0008691), co-funded by the European Union and the European Union Digital Europe project "OpenEuroLLM" (no.: 101195233). The work was further supported by the "Multilingual Lens" project (no.: UNCE/24/SSH/009), and by the SVV project no. 260 821.

References

Ahmed Ahmed, A. Feder Cooper, Sanmi Koyejo, and Percy Liang. 2026. Extracting books from production language models. *Preprint*, arXiv:2601.02671.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, and 84 others. 2025. Apertus: Democratizing open and compliant llms for global language environments. *Preprint*, arXiv:2509.14233.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854, Bangkok, Thailand. Association for Computational Linguistics.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Zikui Cai, Yaoteng Tan, and M. Salman Asif. 2024. Targeted unlearning with single layer unlearning gradient. In *Neurips Safe Generative AI Workshop 2024*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Minseok Choi, Kyunghyun Min, and Jaegul Choo. 2024. Cross-lingual unlearning of selective knowledge in multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10732–10747, Miami, Florida, USA. Association for Computational Linguistics.

Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. 2025. JailbreakRadar: Comprehensive assessment of jailbreak attacks against LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21538–21566, Vienna, Austria. Association for Computational Linguistics.

A. Feder Cooper, Aaron Gokaslan, Amy B. Cyphert, Christopher De Sa, Mark Lemley, Daniel E. Ho, and Percy Liang. 2025. Extracting memorized pieces of (copyrighted) books from open-weight language models. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2025. Everything is editable: Extend knowledge editing to unstructured data in large language models. In *The Thirteenth International Conference on Learning Representations*.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. 2025. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8827–8840, Albuquerque, New Mexico. Association for Computational Linguistics.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, J Zico Kolter, Zachary Chase Lipton, and Pratyush Maini. 2025. Openunlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *The Impact of Memorization on Trustworthy Foundation Models: ICML 2025 Workshop*.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? In *First Conference on Language Modeling*.

Cynthia Dwork. 2006. Differential privacy. In *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *Preprint*, arXiv:2310.02238.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. 2025. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Alphaedit: Null-space constrained model editing for language models. In *The Thirteenth International Conference on Learning Representations*.

Alireza Dehghanpour Farashah, Aditi Khandelwal, Marylou Fauchard, Zhuan Shi, Negar Rostamzadeh, and Golnoosh Farnadi. 2026. Multilingual amnesia: On the transferability of unlearning in multilingual llms. *Preprint*, arXiv:2601.05641.

Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. 2025. On large language model continual unlearning. In *The Thirteenth International Conference on Learning Representations*.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Mingyang Wan, Guojun Ma, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Anyedit: Edit any knowledge encoded in language models. In *Forty-second International Conference on Machine Learning*.

Yongwoo Kim, Sungmin Cha, and Donghyun Kim. 2025. Are we truly forgetting? a critical re-examination of machine unlearning evaluation protocols. *Preprint*, arXiv:2503.06991.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

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.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, and 38 others. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *Preprint*, arXiv:2403.03218.

Taoran Li, Varun Chandrasekaran, and Zhiyuan Yu. 2026. Layer-targeted multilingual knowledge erasure in large language models. *Preprint*, arXiv:2602.22562.

Zexi Li, Xiangzhu Wang, William F. Shen, Meghdad Kurmanji, Xinchi Qiu, Dongqi Cai, Chao Wu, and Nicholas D. Lane. 2025. Editing as unlearning: Are knowledge editing methods strong baselines for large language model unlearning? In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pages 243–254. PMLR.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. 2024. Eight methods to evaluate robust unlearning in llms. *CoRR*, abs/2402.16835.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *Preprint*, arXiv:2401.06121.

Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. *Preprint*, arXiv:2208.03274.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62. Proceedings of the Second EuroHPC user day.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2024. A survey of machine unlearning. *Preprint*, arXiv:2209.02299.

OpenAI. Multilingual massive multitask language understanding (mmmlu). Accessed: Mar. 11, 2026.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. 2025a. Lume: Llm unlearning with multitask evaluations. *Preprint*, arXiv:2502.15097.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. 2025b. Semeval-2025 task 4: Unlearning sensitive content from large language models. *Preprint*, arXiv:2504.02883.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way evaluation for language models. *Preprint*, arXiv:2407.06460.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yue Wang, Qizhou Wang, Feng Liu, Wei Huang, Yali Du, Xiaojiang Du, and Bo Han. 2025. GRU: Mitigating the trade-off between unlearning and retention for LLMs. In *Forty-second International Conference on Machine Learning*.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617, Abu Dhabi, UAE. Association for Computational Linguistics.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

Tianrong Zhang, Bochuan Cao, Yuanpu Cao, Lu Lin, Prasenjit Mitra, and Jinghui Chen. 2025a. WordGame: Efficient & effective LLM jailbreak via simultaneous obfuscation in query and response. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4794–4822, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. 2025b. Catastrophic failure of LLM unlearning via quantization. In *The Thirteenth International Conference on Learning Representations*.