Artificial Intelligence Review (2025) 58:399
https://doi.org/10.1007/s10462-025-11376-7

A survey on large language models unlearning: taxonomy, evaluations, and future directions

Uyen N. Le-Khac[1] · **Vinh N. X. Truong**[1]

Received: 3 October 2024 / Accepted: 31 August 2025

Abstract
Following the introduction of data privacy regulations and "the right to be forgotten", large language models (LLMs) unlearning has emerged as a promising data removal solution for compliance purposes, while also facilitating a diverse range of applications, including copyright protection, model detoxification and correction, and jailbreaking defence. In this survey, we present the taxonomy of existing LLMs unlearning algorithms, summarise unlearning evaluation methods including specialised benchmarks and threat models, and explore the applications of unlearning to provide a broad overview of the current state-of-the-art. We propose a novel problem formulation of LLMs unlearning with the additional unlearning objective: "robustness" to reflect the growing research interest in not only effectively and efficiently eliminating unwanted data, but also ensuring the process is performed safely and securely. To the best of our knowledge, we are the first to examine the robustness of unlearning algorithms as well as threat models for robustness evaluation, aspects that have not been assessed in past surveys. We also identify the limitations of the current approaches, including limited applicability to black-box models, vulnerability to adversarial attacks and knowledge leakage, and inefficiency, all of which require further improvement in future works. Furthermore, our survey highlights future directions for LLMs unlearning research, such as the development of comprehensive evaluation benchmarks, the movement towards robust unlearning and explainable AI for unlearning mechanisms, and addressing potential ethical dilemmas in unlearning governance.

Keywords Large language models · Unlearning · Data deletion · Data privacy · Model detoxification · Robust unlearning

Abbreviations

AI	Artificial Intelligence

✉ Uyen N. Le-Khac
uyen.le-khac@rmit.edu.vn

Vinh N. X. Truong
vinh.truongnguyenxuan@rmit.edu.vn

[1] School of Science, Engineering and Technology, RMIT University, Ho Chi Minh City, Vietnam

Published online: 29 October 2025

ASR	Attack Success Rate
BERT	Bidirectional Encoder Representation from Transformer
CCPA	California Consumer Privacy Act
DUA	Dynamic Unlearning Attack
EUL	Efficient Unlearning Method for LLMs
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformers
ICUL	In-Context Unlearning
IEEE	Institute of Electrical and Electronics Engineers
KGA	Knowledge Gap Alignment
KL	Kullback–Leibler
KnowUnDo	Knowledge Unlearning with Differentiated Scope
LLMs	Large Language Models
LAU	Latent Adversarial Unlearning
LoRA	Low-Rank Adaptation
MIA	Membership Inference Attack
MMLU	Massive Mutiltask Language Understanding
NLP	Natural Language Processing
NPO	Negative Preference Optimisation
PaLM	Pathways Language Model
PIPA	Personal Information Protection Act
PO	Preference Optimisation
RLHF	Reinforcement Learning from Human Feedback
RKLD	Reversed Kullback–Leibler-divergence-based knowledge distillation
SHAP	SHapley Additive exPlanations
SPUL	Soft Prompting for Unlearning
SPUNGE	Split, Unlearn, Merge
SSU	Stable Sequential Unlearning
TOFU	Task of Fictitious Unlearning
ULD	Unlearning from Logit Difference
WMDP	Weapons of Mass Destruction Proxy
XAI	Explainable Artificial Intelligence

1 Introduction

The breakthroughs in large language models (LLMs) research have driven an unprecedented interest in developing and deploying generative models in multiple disciplines, ranging from healthcare, finance, and legal to science and education (Chen et al. 2024a; Chang et al. 2024; Ferdaus et al. 2024; Naveed et al. 2023). As LLMs advance rapidly, critical privacy concerns have been raised (Naveed et al. 2023; Majeed and Hwang 2024; Das et al. 2024), mostly due to the vast public online data LLMs were pre-trained on. While contributing to LLMs' remarkable natural language processing (NLP) capabilities, these web-crawled data may contain private information and copyrighted materials. Including such data carries the risks of privacy violation and copyright infringement (Singhal et al. 2023; Qu et al. 2024; Yao et al. 2024b; Xu 2024; Liu et al. 2024d). Several data privacy regulations have

been introduced in the last decade to address the highlighted issues, making data protection mandatory by law. One of the most foundational regulations is the European Union General Data Protection Regulation (GDPR) (European Union 2016), which introduces new legal concepts such as the "right to be forgotten" and "right to withdraw consent". These legal rights allow individuals to request the removal of their private data from the Internet and databases, and revoke their consent to data controllers. The establishment of GDPR was followed by various other regulations such as the California Consumer Privacy Act (CCPA) (State of California Department of Justice 2018), or South Korea's Personal Information Protection Act (PIPA) (Personal Information Protection Commission 2020).

Several high-profile legal cases surrounding copyright and personal data in artificial intelligence (AI) models have attracted attention from the general public, AI practitioners, and the research community. In 2021, a court decision by the Federal Trade Commission mandated a photo storage application company to delete not only the requested private images but also the facial recognition algorithms and models developed using these images (Federal Trade Commission 2021). In 2023, The New York Times filed a copyright infringement lawsuit against OpenAI for using the newspapers' copyrighted contents to train the LLMs that powered the ChatGPT chatbot (Grynbaum and Mac 2023). The legal landscape indicates a positive transition into data and privacy protection, especially when data are being treated as a market commodity (Custers and Malgieri 2022). However, it also highlights how these regulations were not introduced with generative AI and LLMs in mind, and remain ambiguous (Liu 2024). In the case of LLMs, the problem expands beyond a naive data removal task due to the models' massive size and complex deep learning architecture. As the development of LLMs is computationally expensive, deleting such models due to privacy or copyright violation would be a major setback while retraining them from scratch would be equally costly and impractical (Si et al. 2023; Liu et al. 2024b; Yao et al. 2024a). The presented problems and challenges emphasise an urgent demand for an effective and efficient data erasure solution specifically for LLMs.

Furthermore, as LLMs evolve and are widely integrated into various systems across multiple domains, regulation compliance or copyright protection are no longer the sole motivations for data erasure. Recent years have witnessed the emerging movement towards safe and trustworthy AI, which aims at overcoming the challenges and limitations of LLMs' performance such as harmful content generation, toxic behaviours, societal bias and hallucinating responses (Bender et al. 2021; Wen et al. 2023; Kotek et al. 2023; Yao et al. 2023b; Li et al. 2024b). Inspired by the "machine unlearning" concept first proposed by Cao and Yang (2015) for statistical classification models, various studies have explored the unlearning mechanism in LLMs. In this survey, we focus on the unlearning algorithms designed for LLMs, a new paradigm referred to as "LLMs unlearning".

LLMs unlearning differs from conventional machine unlearning by various factors, including the driven motivations and objectives (Liu et al. 2024b). For machine unlearning, the initial goal is to remove unwanted data from the training data, namely private data as requested, without retraining the model from scratch or compromising the model performance on retaining data (Nguyen et al. 2022; Xu et al. 2024a). The focus on data removal also facilitates the removal of any other undesirable data such as outdated or poisoned data which enhances the model security and trustworthiness (Nguyen et al. 2022). This underlying motivation directs machine unlearning towards two main routes: exact unlearning and approximate unlearning. Exact unlearning introduces strategies to optimise the naive

retraining process for cost reduction (Cao and Yang 2015; Bourtoule et al. 2021; Yan et al. 2022) while approximate unlearning only "approximately" mimics the behaviour of the retrained model, not strictly removing the unlearning data points (Guo et al. 2019; Ullah et al. 2021; Xu et al. 2024a). In light of LLMs, exact unlearning is highly impractical, while some approximate unlearning techniques designed for machine learning models cannot be deployed due to the differences in model architectures (Si et al. 2023). Motivation-wise, LLMs unlearning is applied to a broader set of applications besides data removal, including model detoxification (Yao et al. 2023a; Dige et al. 2024; Lu et al. 2024a; Kadhe et al. 2024; Sheshadri et al. 2024; Li et al. 2024b) and jailbreaking defence (Lu et al. 2024b; Zhang et al. 2024c). These applications result in more diverse unlearning targets of data points and model behaviours and require generalisation on unseen data (Liu et al. 2024d, b). The goal of correcting LLMs behaviours leads to the exploration of alignment-inspired methods such as the variants of preference optimisation algorithms (Rafailov et al. 2024; Zhang et al. 2024a). The black-box setting of various commercial LLMs has also motivated researchers to develop unlearning techniques that require no access to the model parameters such as in-context unlearning (Pawelczyk et al. 2023) and soft prompting (Bhaila et al. 2024). However, the effectiveness of input modification algorithms remains controversial due to the challenges in evaluation and verification. Furthermore, in-context unlearning and soft prompting require storing unwanted data for prompting construction, contradicting the data privacy established by regulations.

1.1 Comparison to related surveys

Several surveys have been carried out to provide insights on LLMs unlearning. Si et al. (2023) were among the first attempts to investigate LLMs unlearning literature, formulate the objectives and develop methods taxonomy. However, the survey only provided a brief overview of the unlearning framework and focused primarily on the effectiveness and utility preservation objectives. Furthermore, it lacks the discussion on evaluation benchmarks and threat models, most of which did not exist at the time of publication.

Similarly, the short survey by Xu (2024) compared LLMs unlearning with traditional machine unlearning, without examining in-depth the evaluation process and potential applications. Liu et al. (2024b) provided a more critical analysis of the LLMs unlearning framework, evaluation methods and applications. The paper, however, did not provide algorithms taxonomy while also not exploring benchmarks and attack-based evaluations in great detail. Blanco-Justicia et al. (2024) was the first comprehensive survey with intricate taxonomy and benchmarks evaluation, yet did not assess the applications and use cases of LLMs unlearning. A recent work by Liu et al. (2024d) on unlearning for generative AI provided a more contemporary overview of the research topic.

Different to past surveys, this study considers the robustness objective of LLMs unlearning. According to the Institute of Electrical and Electronics Engineers (IEEE) glossary, robustness is the degree to which a system can perform accurately under invalid inputs or stressful conditions (IEEE 1990). In the context of LLMs unlearning, this can be further specified as the model ability to maintain its intended functionality in the presence of adversarial attacks and jailbreaking attempts. In particular, the unlearned data should not resurge, and the unlearning process should not introduce new vulnerabilities to the model. The application of jailbreaking defence, which was previously overlooked, is also explored.

Furthermore, we examined the utilisation of threat models to evaluate LLMs unlearning algorithms effectiveness, utility preservation, and also robustness. While previous surveys such as (Liu et al. 2024d; Blanco-Justicia et al. 2024) have explored membership inference attacks (MIA) mechanisms, we examined a more diverse set of threat models, serving a wider set of objective evaluation and verification. A summary of the comparison between our survey and related surveys is presented in Table 1.

1.2 Contributions of this survey

LLMs unlearning is capable of facilitating a wide range of applications and unlearning targets yet remains an under-explored research topic (Liu et al. 2024b). Here, we conduct a comprehensive review of the current state-of-the-art landscape of LLMs unlearning algorithms, frameworks, and evaluation methods. In light of recent research progression, we redefine LLMs unlearning objectives and propose a novel problem formulation to indicate the movement towards robust unlearning. We aim to provide an insightful and up-to-date picture of the algorithms' taxonomy, benchmarks and threat models for evaluation, and use cases of LLMs unlearning. With this survey, we seek to inform and motivate readers to carry out innovative and impactful endeavours in LLMs unlearning, tackle presented challenges and limitations, and advance towards safe, trustworthy and ethical AI.

The main contributions of this survey are listed as follows:

- We propose a novel problem formulation for LLMs unlearning with a new objective introduced, as detailed in Sect. 4.
- We systematically categorise and assess the existing LLMs unlearning algorithms per our proposed objectives, as detailed in Sect. 5.
- We summarise the evaluation methods for LLMs unlearning, including benchmarks and threat models, provided in Sect. 6.
- We survey the applications of LLMs unlearning, including the post-hoc defence role which has not been examined previously, discussed in Sect. 7.

Table 1 A comparison between our survey and existing LLMs unlearning surveys concerning problem formulations, algorithms taxonomy, evaluation methods and applications

Surveys	Objectives				Algorithms		Evaluations			Applications			
	Effectiveness	Efficiency	Utility	Robustness	Taxonomy	Robust unlearning	Computation	Benchmarks	Threat models	Regulation compliance	Copyright protection	Model detoxification	Jailbreaking defence
Si et al. (2023)	●	✗	●	✗	◐	✗	✗	✗	✗	✗	✗	✗	✗
Xu (2024)	●	●	✗	✗	◐	✗	●	✗	✗	✗	✗	✗	✗
Liu et al. (2024b)	●	●	●	✗	✗	✗	●	○	✗	●	●	●	✗
Blanco-Justicia et al. (2024)	●	●	●	✗	●	✗	●	○	✗	✗	✗	✗	✗
Liu et al. (2024d)	●	●	●	✗	●	✗	✗	○	✗	●	●	●	✗
Ours	●	●	●	●	●	●	●	●	●	●	●	●	●

● = in-depth coverage; ◐ = moderate coverage; ○ = brief overview; ✗ = no coverage

- We comprehensively analyse the research landscape and discuss potential directions for future research, as detailed in Sect. 8.

The organisation of the remainder of this paper is as follows. Section 2 presents the methodology of this survey, including search strategy and screening process. Section 3 offers the preliminary background concepts of this survey. Section 4 presents our novel problem formulation of LLMs unlearning. Section 5 provides the taxonomy of existing LLMs unlearning algorithms, their technical foundations and limitations. Section 6 examines the unlearning evaluation methods including computation analysis, benchmarks and threat models, while Sect. 7 explores the applications of LLMs unlearning. Section 8 provides a comprehensive discussion of the current landscape, and highlights potential future directions. Finally, Sect. 9 concludes the survey.

2 Methodology

In this survey, we collected data from two databases: Scopus and the preprint repository arXiv. Due to the rapidly evolving nature of LLMs unlearning research where most publications are in the early stages or published in non-fully peer-reviewed channels, we consider all types of literature on the topic including pre-prints and conference proceedings.

The search queries were formulated to extract relevant literature specifically focuses on LLMs unlearning from 2022 to 2024:

- Scopus: TITLE-ABS-KEY ((“large language model*” OR “LLM*”) AND “unlearn*”) AND (LIMIT-TO (PUBYEAR, 2022) OR LIMIT-TO (PUBYEAR, 2024))
- arxiv: date_range: from 2022-01-01; include_cross_list: True; terms: AND title=large language model*; AND title=unlearn*; OR title=LLM*; AND title=unlearn*

The initial search in September 2024 returned a total of 109 records, comprising 11 articles from Scopus and 98 articles from arXiv. After removing duplicates, 103 articles remained. These papers were manually reviewed by the authors for quality check and empirical evidence. During the reviewing and screening process, we adhere to the inclusion exclusion criteria presented in Table 2. The final number of surveyed articles was 56. Figure 1 illustrates the data collection and screening process.

Table 2 Inclusion and exclusion criteria

Criterion	Inclusion	Exclusion
Literature type	Peer-reviewed journals, books and books chapters, conference proceedings, pre-prints, blogs	Not applicable
Content	Articles specifically focus on LLMs unlearning	Technical papers, discussion papers without methodology, missing abstracts, unrelated to LLMs
Language	English	Non-English
Timeline	2022–2024	Before 2022

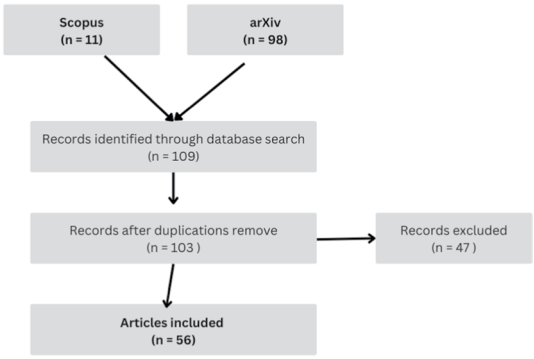

Fig. 1 The data collection process where a total of 109 records were retrieved from Scopus and arXiv. Following duplication removal and manual screening, 56 articles were included in the study

3 Preliminaries

In this section, we present the foundational concepts of LLMs and machine unlearning, which serve as the foundational background for LLMs unlearning development. We also introduce key concepts relevant to LLMs unlearning, including differential privacy, MIA, and adversarial training. Additionally, we discuss related techniques such as knowledge editing and reinforcement learning from human feedback (RLHF), as understanding these methods is essential for identifying the unique objectives and unlearning targets in LLMs unlearning.

3.1 Large language models

LLMs are advanced, state-of-the-art deep learning systems with remarkable capabilities to understand and coherently generate text (Naveed et al. 2023). LLMs can perform a wide range of downstream tasks including content generation, conversational interaction and language translation (y Arcas 2022). LLMs also possess a high level of contextual awareness, and the ability to leverage inputs and follow instructions (Naveed et al. 2023; Yang et al. 2024).

Building upon the self-attention mechanism of Transformers architecture, the history of LLMs can be traced back to the early works such as Google's Bidirectional Encoder Representation from Transformers (BERT) (Devlin 2018) and OpenAI's Generative Pre-trained Transformers (GPT) (Radford et al. 2018). Most LLMs were pre-trained on massive text for general text representation with the capability of generalisation to diverse sets of unseen tasks (Sanh et al. 2021). However, they can also be fine-tuned for specific downstream tasks, domain adaptations or human preference alignment using a small set of instruction tuning data (Liu et al. 2024a, c; Li et al. 2024a), enabling their widespread adoption in various tasks and domains.

A significant characteristic of LLMs is the large number of parameters and the massive training text corpora involved in their pre-training phase (Yao et al. 2024b). Therefore, developing LLMs is computationally expensive with extended training time overhead (Acharya et al. 2023; Fields et al. 2024). As LLMs evolve, the number of parameters continuously increases, going from 342 million in the BERT models (Devlin 2018), to 175 billion in the GPT-3 model (Brown 2020), to 540 billion in Pathways Language Model (PaLM) (Chowdhery et al. 2023), making these model development even more resource-intensive.

However, these numbers have yet to accurately represent LLMs' evolution as many recently released models with more sophisticated performances such as the GPT-4 (Achiam et al. 2023) are black-box models where the parameter counts remain unpublished.

3.2 Machine unlearning

Machine unlearning was first introduced by Cao and Yang (2015) to perform data forgetting for statistical classification models. The machine unlearning process aims to remove specified learned data and associated knowledge from the learned model, reversing the learning process conducted on unwanted data (Shaik et al. 2023). An effective unlearning algorithm will result in an unlearned model that performs as if it has never learned from the unlearned data samples (Xu et al. 2024a). Past studies revealed the evolution of machine unlearning, where the research interest shifted from the golden-standard exact unlearning to the more light-weight and efficient approximate unlearning (Yan et al. 2022; Liu et al. 2024b).

3.2.1 Exact unlearning

A naive method for data removal is retraining the model from scratch after deleting the undesirable data, as defined by Cao and Yang (2015) in Definition 1. This approach guarantees the complete elimination of such data from the model (Nguyen et al. 2022) but remains computationally expensive and impractical (Thudi et al. 2022). Furthermore, it cannot be applied in cases where training data is inaccessible, for instance, federated learning (Gong et al. 2022).

Definition 1 [Naive Retraining (Cao and Yang 2015)] Given the learning algorithm A(.), training set D, forget set D_f, retraining process R(.), the parameters of naive retrained model w_r is:

$$w_r = A(D \setminus D_f)$$

Due to the limitations and impracticality of naive retraining, exact unlearning was proposed in the machine unlearning pioneering works. While also involves retraining the model, exact unlearning employs training strategies to optimise the retraining process to reduce computational cost and complexity. The formal definition of exact unlearning provided by Nguyen et al. (2022) is presented in Definition 2, in which the objective of exact unlearning is to ensure that the distribution of the unlearned model is consistent and indistinguishable from the retrained model.

Definition 2 [Exact Unlearning (Nguyen et al. 2022)] Given the learning algorithm A(.), training set D, forget set D_f, and unlearning algorithm U(.), exact unlearning can be defined as:

$$Pr(A(D \setminus D_f)) = Pr(U(D, D_f, A(D)))$$

Most of the exact unlearning studies suggested strategies to localise and limit the retraining samples or affected sections of models that require retraining to improve efficiency. Cao and

Yang (2015) suggested converting the learning algorithm into summation form. Therefore, when the specified data points are removed, only a limited number of summations will be affected. Bourtoule et al. (2021) introduced the Sharded, Isolated, Sliced, and Aggregated (SISA) training strategy which divides training data into disjoint data shards and develops corresponding sub-models. Similarly, retraining when enacted will only perform updates on a small group of shards and sub-models, speeding up the process. Similarly, the study An Efficient Architecture for Exact Machine Unlearning (ARCANE) utilised one-class classifiers and novel data pre-processing techniques to reduce the number of retraining samples, making the retraining time faster (Yan et al. 2022).

3.2.2 Approximate unlearning

As exact unlearning requires significant computational power and is only feasible for simple models (Xu et al. 2024a), researchers have developed approximate unlearning which is more scalable and less resource-intensive (Liu et al. 2024b). Unlike exact unlearning, approximate unlearning does not specifically remove the D_f from the training data but instead, mimics the performance of a model which have not learned from these data. The objective of approximate unlearning is to ensure the unlearned and retrained models' distributions remain approximately indistinguishable (Nguyen et al. 2022), instead of a guarantee similarity in exact unlearning. The common approach to guarantee the distribution approximation is by utilising the differential privacy concept which inspired the foundation of the probabilistic notion of unlearning objectives (Guo et al. 2019; Ullah et al. 2021), as presented in Definition 3.

Definition 3 [(ϵ, δ)- Approximate Unlearning (Guo et al. 2019)]

Given $\epsilon, \delta > 0$, unlearning algorithm U(.) achieves ϵ-certified removal for learning algorithm A(.) if $\forall \tau \subseteq H, D \in Z^*, z \in D$:

$$Pr(U(D, z, A(D)) \in \tau) \leq e^{\epsilon} Pr(A(D \setminus z) \in \tau) + \delta$$

and

$$Pr(A(D \setminus z) \in \tau) \leq e^{\epsilon} Pr(U(D, z, A(D)) \in \tau) + \delta$$

According to Xu et al. (2024a), approximation unlearning falls into four main categories: influence function, re-optimisation, gradient update, and graph methods. Influence function-based methods estimate the influence of data points by computing or approximating the influence function, then update the model parameters to remove this influence. The re-optimisation approach involves iteratively updating the model parameters to remove the influence of forget data points while preserving the model utility by minimising the loss function on the retain data. On the other hand, gradient-based technique performs gradient updates to modify model parameters towards forgetting specific data points. Lastly, graph unlearning methods are designed specifically for graph neural networks to eliminate unwanted graph-structured data.

3.3 Differential privacy

Differential privacy is a privacy guarantee method that applies an additive noise mechanism on training data to obtain the certificates of privacy (Dwork et al. 2006, 2014). The application of differential privacy will ensure that for any given data points in the training data, their influence and impact on the model output are minimal, making the inclusion or exclusion of any data remain undetected (Chen et al. 2021). In other words, differential privacy makes it challenging to retrace the training data samples via model output examination, and therefore prevent MIA (Jayaraman and Evans 2019; Chen et al. 2021). In the context of machine unlearning, "differential privacy implies approximate unlearning" (Nguyen et al. 2022), and DP shares similar goals in data privacy protection as unlearning algorithms. However, it is noted that differential privacy often comes with a high loss in the model's accuracy, even when the initialised ϵ is large and only a weak privacy guarantee is secured (Chaudhuri et al. 2011; Abadi et al. 2016).

3.4 Membership inference attacks

MIA, as formulated by Shokri et al. (2017) for machine learning problems, are commonly used as attack-based methods to determine whether a given data point was included in the training dataset of a trained model. A key characteristic of MIA is that it does not require access to the model architecture or the underlying distribution of training data. Instead, an attack model can be trained using a set of shadow models that mimic the behavior of the target model. Then, using the attack model, attackers can acquire labels to classify whether a data point is a member of the training set, effectively revealing membership information (Shokri et al. 2017). For LLMs, MIA serves a dual role as both a potential threat and a valuable evaluation tool for unlearning algorithms. On the one hand, MIA can be used to extract knowledge from LLMs, even after an unlearning process has been applied, highlighting potential vulnerabilities. On the other hand, when employed for evaluation, MIA acts as a key metric to assess the robustness and effectiveness of unlearning algorithms (Blanco-Justicia et al. 2024), especially since retraining for evaluation is not an option (Liu et al. 2024b).

3.5 Adversarial training

Adversarial training, as per Definition 4, was proposed by Madry et al. (2017) as a novel approach to enhance deep learning model robustness against adversarial inputs. The technique is considered to be a defense technique against adversarial attacks via optimisation problem formulation (Silva and Najafirad 2020).

Definition 4 [Adversarial Training (Madry et al. 2017)] Given input x and corresponding label y, model parameters θ, data distribution D, loss function L, perturbation δ, and population risk $\mathbb{E}_D[L]$, adversarial training is formulated as a saddle point minimisation problem:

$$\min_{\theta} p(\theta), where \;\; p(\theta) = \mathbb{E}_{(x,y)\sim D}[\max_{\delta \in S} \;\; L(\theta, x+\delta, y)]$$

The approach is formulated as a saddle point optimisation problem which involves an interlink inner maximisation and outer minimisation function. The inner maximisation aims at maximising the adversarial loss on given input data, while the outer minimisation optimises the model parameters in the direction that minimises the inner attack loss. This interlink formulation encapsulates both the attacking and defending mechanisms, enabling a high guarantee against adversarial attacks (Madry et al. 2017).

3.6 Knowledge editing

Knowledge editing emerges as a solution to update and correct LLMs' output as the world's state of information progresses and evolves (De Cao et al. 2021). The process locally modifies specific knowledge within the knowledge base of LLMs without influencing the retaining knowledge. Knowledge editing seeks to improve the model performance and consistency without retraining or fine-tuning the entire architecture (Sinitsin et al. 2020; De Cao et al. 2021; Wang et al. 2023b; Yao et al. 2023a). When applied, knowledge editing will perform model manipulation by updating a large number of parameters in the model to enable the modification process (De Cao et al. 2021; Mitchell et al. 2021; Meng et al. 2022; Hase et al. 2023).

There are two mechanisms by which knowledge editing can be applied to LLMs: (1) knowledge insertion, and (2) knowledge modification. Knowledge insertion allows new pieces of knowledge to be injected into the model as they emerge to keep the model up-to-date and reliable (Martino et al. 2023). On the other hand, knowledge modification facilitates the correction and update of specific knowledge within the knowledge base of the LLMs (Song et al. 2024).

3.7 Reinforcement learning from human feedback

RLHF was first introduced by Christiano et al. (2017) as a novel approach to solving complex reinforcement learning (RL) problems without having to access the model's reward function, by using only a small set of human preference feedback. As RLHF involve training a reward function using human feedback instead of using the feedback directly as a reward system, it is more sample-efficient and less labour-intensive (Christiano et al. 2017). The underlying mechanism of RLHF is to reward and encourage the model's desirable behaviours that align with human preferences and objectives. RLHF requires a collection of human feedback in the form of prompt and response pairs (Christiano et al. 2017). The feedback often follows the binary rating system (Li et al. 2016; Scheurer et al. 2023), or the ranking system where human experts pick out the most appropriate and preferable response among a set of instructions (Ziegler et al. 2019), and can be manually created or automatically generated using LLMs (Chaudhari et al. 2024).

For LLMs, RLHF is a mainstream technique to perform model alignment, commonly via red teaming or user reporting (Yao et al. 2023b). Askell et al. (2021) defined the objectives of RLHF integration in LLMs as (1) helpfulness, (2) honesty, and (3) harmlessness. These goals imply that LLMs should generate informative, relevant, and trustworthy responses while refraining from providing harmful, biased and sensitive knowledge. OpenAI has leveraged RLHF to align the responses of ChatGPT with human preference and optimise the chatbot's conversational ability and trustworthiness (OpenAI 2024).

4 Large language models unlearning

4.1 Problem formulation

LLMs unlearning is built upon conventional machine unlearning, originally established by Cao and Yang (2015), while presenting a unique set of challenges. First, unlike conventional machine unlearning, which is designed for smaller and less complex models such as statistical classification algorithms, LLMs unlearning must account for the large scale of training data, model parameters, and the generative nature of LLMs. Second, conventional machine unlearning aims at removing isolated data points, whereas LLMs' forget set can extend to behaviours or concepts with less defined boundaries (Liu et al. 2024b). Third, while machine unlearning follows exact and approximate approaches, as discussed in Sect. 3.2, LLMs' deep learning architectures render exact unlearning impractical and infeasible, and limit the effectiveness of approximate methods due to architectural adaptability and computational efficiency concerns, given the large number of parameters involved (Bucknall and Trager 2023).

One of the early works by Si et al. (2023) defined LLMs unlearning objectives to be effectiveness and locality (utility preservation). Liu et al. (2024b) further added the efficiency objective for unlearning algorithms, in which they should be highly efficient in comparison to the naive retraining approach. Similarly, Blanco-Justicia et al. (2024) also focused on three main objectives: effectiveness, efficiency, and utility. However, in light of recent works in LLMs unlearning, we observe a new research interest focusing on the robustness of unlearning algorithms (Zhang et al. 2024c; Lu et al. 2024b; Sheshadri et al. 2024; Yuan et al. 2024). These studies targeted not only the three established objectives but also the robustness capability of LLMs unlearning algorithms. The emergence of a new research focus has motivated us to re-define the problem formulation for LLMs unlearning with the following four objectives:

1. *Effectiveness*: The effectiveness in eliminating unlearning targets influence and any associated capabilities on a learned model. Ideally, the unlearned model should behave as if it has never been trained on unlearned data.
2. *Efficiency*: The requirement of runtime and computational resources of the unlearning process in comparison to the naive retraining approach.
3. *Utility*: The performance on retaining data and general model utility should be preserved following the unlearning process.
4. *Robustness*: The degree to which the unlearned model can maintain its intended functionality in the presence of adversarial inputs, jailbreaking attempts, or other exploitation techniques. Ideally, the unlearning process should not introduce new vulnerabilities or decrease the model security under threatening conditions.

4.2 Association with related concepts

LLMs unlearning is closely related to several concepts such as knowledge editing and RLHF. The concept of model alignment and RLHF inspires the development and application of LLMs unlearning for model detoxification purposes. Meanwhile, knowledge editing and LLMs unlearning share a common goal of knowledge base modification. However, the

techniques are distinctly separated by the objectives and task definition. Table 3 summarises the key differences between LLMs unlearning and the related concepts.

Despite having an insignificant connection to machine unlearning, RLHF (as defined in Sect. 3.7), is closely related to LLMs unlearning. Due to the expanded applications of LLMs unlearning, the goals of LLMs partially resemble model alignment methods such as RLHF. However, instead of encouraging the model desirable behaviour like alignment technique, LLMs unlearning aims at not exhibiting undesirable behaviours. In other words, instead of giving desirable responses, LLMs unlearning focus on not providing undesirable answers (Yao et al. 2023b). Additionally, RLHF primarily targets aligning the model output with human preference, while LLMs unlearning is a broader concept. In essence, RLHF leans towards model alignment, while LLMs unlearning can be used for model detoxification, jailbreaking defence and data removal tasks. Another key difference between the two concepts is the required samples: unlearning only needs negative samples, whereas RLHF requires a more comprehensive and complex set of human preference data which includes positive and negative samples.

For knowledge editing, as discussed in Sect. 3.6, the technique focuses on modifying knowledge locally, whereas unlearning seeks to forget specific unwanted data or divert away from undesirable behaviours. In the context of LLMs, several unlearning algorithms are capable of generalising on unseen data, and eliminating data with similar unwanted characteristics.

5 Taxonomy of large language models unlearning algorithms

In this section, we examine the unlearning algorithms specifically designed for LLMs and classify them into three main categories: parameter modification, input modification, and robust unlearning. Parameter modification techniques are algorithms that require parameter optimisation or modification to perform unlearning. Occasionally, the algorithms will require modification directly to the model architecture. In contrast, the input modification models are prompt engineering-inspired and do not access or modify the model parameters and architectures. Additionally, we examine a new category of "robust unlearning", which shifts the focus towards robust unlearning on aligned LLMs. The algorithms are assessed and measured based on the four objectives proposed in Sect. 4.1. Due to the large size of data and model parameters, most LLMs unlearning algorithms follow the principles and objectives of approximate unlearning in conventional machine unlearning, in which the model mimics the behaviours of retrained models without exactly removing the unwanted data points. As a result, most LLMs unlearning models achieve "approximate" unlearning effect, except for input modification methods, which provide no guaranteed unlearning due to their nature of a prompt-engineering technique. The taxonomy of unlearning algorithms is visualised in Fig. 2.

Table 3 Comparison LLMs unlearning and related concepts

	LLMs unlearning	Knowledge editing	RLHF
Definition	A technique to eliminate the influence of unwanted data or undesirable behaviours from LLMs	A technique to locally modify existing knowledge or insert emerging knowledge into LLMs	A technique to align the model outputs to human preferences
Applications	Regulation compliance, copyright protection, model detoxification, and jailbreaking defence	Knowledge modification towards trustworthy LLMs	Model alignment to optimise the model responses towards human preferences and intentions
Method	Apply unlearning algorithms to reverse the learning process on unwanted data or steer the model away from undesirable behaviours	Make modifications to the model parameters and knowledge base to edit a specified knowledge	Train a reward function using human feedback and optimise it via reinforcement learning
Targets	Unwanted data or undesirable behaviours. (Negative samples)	Specific knowledge requiring modification or injection	Human feedback corresponds to model responses. (Positive and Negative samples)

5.1 Parameter modification

5.1.1 Gradient-based

Gradient-based algorithms include gradient ascent, gradient descent, gradient difference, and their variants. Due to the straightforward objective function and pioneering characteristic, gradient-based algorithms are one of the most commonly applied approaches in LLMs unlearning, often serve as the baseline models for further development or performance comparison in various studies (Jang et al. 2022; Eldan and Russinovich 2023; Maini et al. 2024; Zhang et al. 2024a; Yao et al. 2024a; Jia et al. 2024; Dou et al. 2024; Bhaila et al. 2024).

Among the gradient-based algorithms, gradient ascent is highly prevalent in LLMs unlearning literature. As defined in Definition 5, the gradient ascent objective function aims to reverse the learning process on the forget set D_f by updating the model parameters towards the direction of increasing loss. The ultimate goal is to maximise the likelihood of inaccurate predictions within D_f (Golatkar et al. 2020).

Definition 5 [Gradient Ascent (Golatkar et al. 2020)] Given the model input x and corresponding label y, forget set D_f, loss function L, and set of parameters θ, gradient ascent aims at:

$$\min_{\theta} \; -\mathbb{E}_{x,y\in D_f}[L(y|x;\theta)]$$

Various studies have utilised vanilla gradient ascent to perform unlearning tasks on pretrained LLMs (Jang et al. 2022; Maini et al. 2024; Gu et al. 2024; Yuan et al. 2024; Lu et al. 2024a). However, it was observed that unlearning via gradient ascent led to catastrophic collapses in the model utility due to the algorithm's excessive unlearning habit (Zhang et al. 2024a; Wang et al. 2024b). To tackle this challenge, gradient ascent is often paired with Kullback–Leibler (KL) divergence or another algorithm to achieve a more balanced unlearning effectiveness and model utility trade-off. KL divergence is a statistical metric to measure the distance between two model distributions. In the context of machine unlearning, Golatkar et al. (2020) formulated a minimisation problem on KL divergence to achieve the unlearning goals. The combination of gradient ascent and KL divergence minimisation allows the model to minimise the differences in distribution between unlearned and retrained models while maximising the loss on the forget set. The approach was also taken by Yao et al. (2023b) for LLMs unlearning, while Yao et al. (2024a) paired gradient ascent with gradient descent to improve the robustness of the hyperparameters.

Another variant of gradient ascent is gradient difference, which adds a regularising term to the formulation to preserve the model capabilities on non-target data points D_r (Liu et al. 2022). As seen in Definition 6, gradient difference inherits gradient ascent's goal to maximise the loss on the forget set D_f, but also include the condition to minimise the loss on the retain set D_r. Generally, gradient ascent is gradient difference when the regularisation parameter λ is set to nil. Similar to gradient ascent, gradient difference was also employed in several LLMs unlearning studies as baseline unlearning frameworks (Maini et al. 2024; Jia et al. 2024).

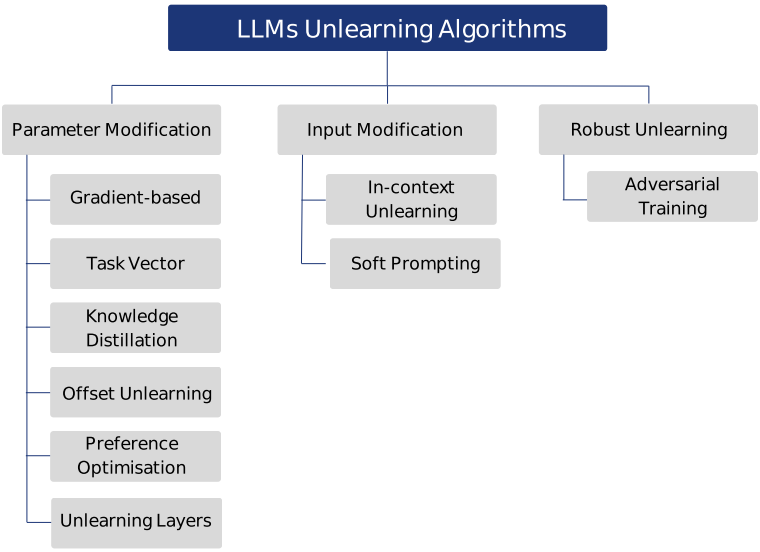

Fig. 2 The taxonomy of existing LLMs unlearning algorithms includes three main categories: parameter modification, input modification, and robust unlearning

Definition 6 [Gradient Difference (Liu et al. 2022)] Given the model input x and corresponding label y, forget set D_f, loss function $\mathcal{L}$, parameter weight θ and regularisation parameter λ, gradient difference performs:

$$\min_{\theta} \; -\mathbb{E}_{x,y\in D_f}[\mathcal{L}(y|x;\theta)] + \lambda \mathbb{E}_{x,y\in D_r}[\mathcal{L}(y|x;\theta)]$$

Gradient descent was also applied in a foundational work of LLMs unlearning by Eldan and Russinovich (2023). The technique consists of three main stages: obtaining a reinforced model, relabelling unlearning data points, and fine-tuning. First, the baseline language model is further fine-tuned on the forget set to develop a reinforced model. Then, by comparing the logits of the reinforced model with the baseline model, the tokens with high probabilities of generating unlearning data-related content are highlighted and marked as the "target tokens". Second, the target tokens are substituted with generic, alternative terms and are subsequently relabelled via model prediction. The new labels approximate the behaviour of a model that has never been trained on the target tokens. Finally, the model is fine-tuned based on the relabelled data, allowing the model to "forget" the original text whenever prompted with target context-related data. In contrast to gradient ascent, the goal of the gradient descent loss function is to minimise the likelihood of accurate predictions on relabelled forget samples D_f. However, this approach has potential limitations. The unlearned model is prone to generate hallucinated responses to cover the unlearned content knowledge. Furthermore, it is considered to be impractical for unlearning sets that do not have unique attributes or pronounced concepts, due to the mechanism of translating context-tokens to generic terms.

Apart from Eldan and Russinovich (2023) work which was computationally expensive due to data translating and re-labelling, most gradient-based methods offer moderate effi-

ciency. Through computational efficiency analysis, Yao et al. (2024a) proved that gradient ascent and its variants achieved 10^5 more efficiency runtime cost compared to naive retraining. Additionally, gradient-based models are robust to MIA and prevent privacy leakage on unlearned data (Yao et al. 2024a; Jia et al. 2024). However, the extent of robustness against adversarial attacks and jailbreaking has not been examined in recent works.

5.1.2 Task vector (task arithmetic)

Building upon the weight interpolation and task arithmetic concept, task vector was introduced by Ilharco et al. (2022) as a behaviour-guiding technique for deep learning models. By definition, a task vector is derived by subtracting the weights of a pre-trained model from its fine-tuned version on a specific task. Findings indicated that performing arithmetic operations on task vectors can effectively manipulate the model behaviours: (1) negating a task vector can reduce the model performance on the corresponding task, (2) merging task vectors can enhance the model multi-tasking ability, and (3) forming an analogical relationship between task vectors can improve the model generalisation ability on unseen tasks. Past studies have employed task vector techniques to perform LLMs unlearning, specifically through the negation and addition mechanism. It is observed that negative task vector methods achieved a balanced trade-off between unlearning effectiveness and model preservation in comparison to gradient ascent models which degraded the model utility significantly (Ilharco et al. 2022; Dige et al. 2024). Task vector unlearning methods have not been assessed under the robustness criteria, and also offer a medium efficiency.

Zhang et al. (2023) explored applying negative task vector in conjunction with the parameter-efficient fine-tuning module to develop the negated-LoRA model which offers diverse skills for domain adaption. Dou et al. (2024) further improved the trade-off balance between unlearning effectiveness and utility through the novel method of Stable Sequential Unlearning (SSU). The technique incorporated the task vector model with additional noise via random labelling loss for stability and weight saliency mapping to reduce the risk of catastrophic collapse, then performed sequential unlearning to update the parameters. However, it is noted that negative vector-based models still inadvertently affected non-target knowledge, reduced the model's reasoning capabilities and were hyperparameters sensitive (Dou et al. 2024; Dige et al. 2024). Task vectors were also employed in the data-driven LLMs unlearning framework "Split, Unlearn, Merge" (SPUNGE) (Kadhe et al. 2024). SPUNGE proposed splitting the unlearning data sets into subsets based on the data attributes and carrying out the unlearning process separately on each subset. Then, the unlearning models will be merged using a variant of task arithmetic called TIES-Merging (Yadav et al. 2024) which facilitates multiple models' parameters merging. As various models are required to be fine-tuned under this framework, SPUNGE remains computationally expensive by nature.

The underlying concept of task vector also inspired Zhou et al. (2023) to develop the "security vectors" which make the harmful data unlearnable and prevent the model from generating harmful contents. The security vectors θ_s are additional model parameters trained with harmful data which encourages the model to exhibit harmful behaviours. Then, the model is fine-tuned with forward propagation where the security vectors are frozen while the remaining parameters θ are optimised. In essence, the introduction of the security vectors ensures that the model predicts consistently with harmful data while steering the model parameters θ update away from the harmful direction. The security vectors θ_s will

then be deactivated during inference, allowing LLMs to perform normally without showing undesirable behaviours. However, the security vector approach is highly sensitive to the learning rate hyperparameters. A high learning rate may result in harmfulness increasing and sabotaging the effectiveness of security vectors. Furthermore, it also makes the model prone to overfitting problems which degrade the model utility significantly.

5.1.3 Knowledge distillation

Knowledge distillation is a knowledge compression technique in machine learning (Hinton 2015). The general idea is to transfer the knowledge of a larger deep neural network (teacher model) to a smaller deep neural network (student model), enabling the student model to mimic the prediction of the teacher model in an efficient and less resource-intensive mechanism. The approach trained a distilled model on a knowledge transfer set, then computed soft target probabilities to capture the knowledge from the teacher model. Knowledge distillation often minimises the KL divergence loss function to match the soft targets of the student model with the teacher model.

Utilising the knowledge distillation concept, Wang et al. (2023a) introduced the Knowledge Gap Alignment (KGA) framework. The study defined the term "knowledge gap" as the distance between the distributions of two models with similar architectures but trained on different data sets. In the case of LLMs unlearning, KGA aims at aligning the knowledge gap between the unlearned and the fine-tuned model, allowing them to have similar performance. To kick-start the process, the output model parameters were first initialised using the original training data set, then updated using KGA on the forget set D_f to achieve unlearning goals. KGA framework also introduced a small set of external data D_n and identified two goals: (1) minimising the output distribution between the unlearned model (student model) and the original model on unseen data D_n (teacher model), and (2) preserving the unlearned model capabilities on non-target data points D_r. KGA also applied the KL divergence metric to measure the distance in the output distribution between the two models. However, KGA is an expensive framework due to two reasons: (1) it requires storage for the additional data set D_n and two models A_n and A_f, and (2) it simultaneously trains two models and performs fine-tuning on the entire set of parameters. Lu et al. (2024b) also employed knowledge distillation as a part of the unlearning objective function to preserve the model's general knowledge when performing on retaining data. Similarly, the distillation objective is adopted to retain the next token prediction with the original post-unlearning model acting as the teacher model. KGA was proven to achieve a defence success against MIA close to exact unlearning (Wang et al. 2023a).

A reversed version of KL-divergence-based knowledge distillation (RKLD) was also proposed by Wang et al. (2024a). The approach leveraged reverse KL divergence as the loss function instead of the mainstream forward version. The forward KL divergence applies penalties when the probability distribution of the teacher model is significantly lower than the student model's, assuring that the important tokens in the teacher model are also granted high probabilities in the student model. On the other hand, the reverse KL divergence will avoid assigning high probabilities with tokens not present in the teacher model. As a result, the objectives of reverse KL divergence are more closely aligned with unlearning goals, emphasising data forgetting and learning avoidance (Wang et al. 2024a). The study, how-

ever, acknowledged the uncertainty in unlearning effectiveness on uncontrolled noisy data, and the side effects prevail on the model in the long run.

5.1.4 Offset unlearning

Instead of the mainstream unlearning algorithms such as gradient ascent or task vector which directly update the model parameters, recent works have established a novel paradigm called "offset unlearning" with the ability to perform *passive* unlearning without modifying the LLMs parameters (Huang et al. 2024; Ji et al. 2024). Offset unlearning perform fine-tuning or model training on offset models that are small in scale, then uses logit difference computation to update the logits of the targeted LLMs. As a result, offset unlearning is highly efficient in comparison to naive retraining or other direct parameter modification unlearning methods (Huang et al. 2024; Ji et al. 2024).

Huang et al. (2024) was the first to propose offset unlearning for LLMs with the δ -UNLEARNING framework. The framework involves an ensemble of models: language model M, and two offset models M_o and M_o' initialised at the checkpoint and resulted in nil offset difference across all data. During the unlearning process, only the parameters of A_o' are updated while the remaining two models remain intact. The logit ensemble is utilised to generate outputs. The core idea behind δ-UNLEARNING is to drive the offset model M_o' from the frozen model M_o when exposed to sensitive queries, then learn the optimal logit difference to effectively steer the model prediction away from giving sensitive information. Due to the nature of offset unlearning problem formulation, δ-UNLEARNING can be used even in black-box scenarios without accessing the model parameters, and achieves high efficiency as it only concerns smaller offset models with fewer parameters for tuning. It is noteworthy to mention that δ-UNLEARNING still demands a minimum white-box setting since it requires full access to the model logit. Additionally, δ-UNLEARNING claims to have a high privacy level as the framework does not store sensitive unlearned data (Huang et al. 2024). However, it has not been formally assessed with MIA and attacks for robustness evaluation.

Similarly, Ji et al. (2024) also leveraged the offset unlearning mechanism and reversed the direction of optimisation in unlearning. The study proposed the Unlearning from Logit Difference (ULD) framework and introduced the concept of assistant models to LLMs unlearning. ULD trains an assistant model to memorise the forget set D_f, then performs unlearning by subtracting the logits of the assistant model from the target LLM. Since the assistant model only needs to learn the forget set which is much smaller in comparison to the entire training data, it is considered to be a less challenging task for a language model. The study also suggested using parameter-efficient fine-tuning algorithms such as Low-Rank Adaptation (LoRA) (Hu et al. 2021) to reduce the number of parameters and consequently, the training time. The study also highlighted ULD's ability to preserve the model utility, as a result of its unique bounded objectives which prevent unbounded forget loss by minimising instead of maximising it, while also avoiding unbounded retaining loss since it drives the output distribution towards uniform distribution. Similar to δ-UNLEARNING, ULD's robustness to adversarial attacks has not been comprehensively evaluated.

5.1.5 Preference optimisation

Inspired by the novel model alignment technique direct preference optimisation (Rafailov et al. 2024), various studies have also attempted utilising the preference optimisation (PO) for LLMs unlearning (Maini et al. 2024; Jia et al. 2024; Sheshadri et al. 2024; Gu et al. 2024). The key difference between PO and gradient ascent is the substitution of gradient ascent's unbounded loss for the alignment-based loss of PO, as depicted in Definition 7. Therefore, PO-based models offer similar efficiency to gradient-based techniques. PO loss is calculated based on the preferred unlearning label y_f on the forget set D_f. Jia et al. (2024) suggested that the response of post-PO should be either reject-based such as "I don't know", or a similar type of avoiding answer. Due to the design of its loss function, preference optimisation models are less likely to encounter catastrophic collapses like gradient-based models. However, they are more vulnerable to MIA and adversarial attacks (Jia et al. 2024).

Definition 7 [Preference Optimisation (Jia et al. 2024)] Given the model input x and corresponding label y, forget set D_f and corresponding label y_f, loss function $\mathcal{L}$, parameter weights θ and regularisation parameter λ, PO is an optimisation problem:

$$\min_{\theta} \ \mathbb{E}_{x,y\in D_f}[\mathcal{L}(y_f|x; w\theta] + \lambda \mathbb{E}_{x,y\in D_r}[\mathcal{L}(y|x;\theta)]$$

Zhang et al. (2024a) further derived the concept of PO to apply only to negative examples and proposed Negative Preference Optimisation (NPO). The unlearning problem is then formulated as a preference optimisation task by minimising the NPO loss function to achieve the unlearning goals. Instead of providing the preferred response y_f, NPO applies the vanilla direct preference optimisation (Rafailov et al. 2024) but only on negative samples. Zhang et al. (2024c) developed the Safe Unlearning framework objective function based on NPO loss, while also employing DPO as the baseline model. NPO has also served as the baseline model in several LLMs unlearning studies (Yuan et al. 2024; Jia et al. 2024; Gu et al. 2024).

5.1.6 Unlearning layers

Apart from parameter optimisation or parameter merging, Chen and Yang (2023) proposed the Efficient Unlearning method for LLMs (EUL), which introduces an additional unlearning layer to the model architecture to facilitate unlearning. The unlearning layers were designed based on teacher-student objectives, in which the KL divergence between the output model and the original model on retain data is minimised while maximising the distance to the forget set. Secondly, the model also minimises the task loss on retain data to maintain the model general capabilities. Lastly, a masked language modelling step is carried out to ensure the security of unlearned data against adversarial attacks. These unlearning layers are fused into LLMs architecture, offering a lightweight and efficient solution for unlearning. However, the study only experimented on smaller backbone LLMs, which left the effectiveness of EUL on large commercial LLMs to be undetermined, as well as the robustness aspect of the framework.

5.2 Input modification

While the parameter modification models follow the approximate unlearning path, several studies have attempted to use prompt engineering approaches to perform unlearning. In this case, we do not require access to the model architecture or parameter weights but instead, directly ask the model to unlearn specific data and knowledge via prompting. Input modification tackles two challenges of the parameter modification methods: (1) black-box applicability, and (2) the algorithm efficiency. However, the effectiveness of current input modification techniques is controversial due to the difficulty in evaluation and verification. Furthermore, these approaches yield no forgetting guarantee and offer no data privacy in the context of "the right to be forgotten" due to its requirement of unwanted data storage.

5.2.1 In-context unlearning

In-context unlearning emerges as the novel approach to perform unlearning without modifying the LLMs parameters. Unlike the model-based methods which update the parameter weights θ, in-context unlearning provides input data in the context that induces the model to behave as if it was retrained on retaining set D_r (Pawelczyk et al. 2023). As a result, in-context is independent of the model architecture and can be applied to both white-box and black-box settings.

Pawelczyk et al. (2023) introduced the In-Context Unlearning (ICUL) framework for LLMs unlearning which eliminated the retraining or fine-tuning process. To perform unlearning in question-answering tasks, ICUL involves three steps: forget answer modification, correct answer addition, and model prediction. First, answers within the forget set D_f are swapped into random answers to remove their influences on the model output. These new answers are then added to the query template. Then, the framework randomly samples and adds correct answers into the template to avoid over-correction on the forget set. Finally, the prompt based on the generated template is used as the input for the model to predict the next token. However, the approach renders several limitations such as the lack of unlearning guarantee and the vulnerability against adversarial attacks. ICUL is also unable to process larger deletion requests without resulting in a significant drop in accuracy and is computationally expensive.

5.2.2 Soft prompting

Another state-of-the-art approach within the input-based category is soft prompting, which also facilitates LLMs unlearning without accessing the model parameters. Similar to in-context unlearning, soft prompting induces data forgetting through input prompts and is independent of the model architecture and training data size. Yet, it does not require manual instruction or sample context prompts but instead, automatically and continuously optimises prompts (Bhaila et al. 2024).

Leveraging the soft prompting concept with specific unlearning objectives identified, Bhaila et al. (2024) proposed the Soft Prompting for Unlearning (SPUL) framework for LLMs. The methodology of SPUL entails three main objectives, formulated as three loss functions. For the first objective, SPUL encourages data forgetting on the forget set D_f by forcing the model to associate the samples within the forget set with an alternative generic

label instead of the actual label. The second objective is maintaining model utility on retaining data, which means the prompt tokens must not change the predictive sequence within the retaining dataset. The third objective of SPUL is to ensure that the unlearned model does not deviate far from the original model using the concept of KL divergence.

Bhaila et al. (2024) pointed out that SPUL adapted better to unlearning tasks with larger LLMs. The size of the forget set does not affect utility preservation but a larger forget set tends to result in better unlearning effectiveness. However, it is noted that the approach was not extensively evaluated and verified, and the risks of adversarial attacks and information leakage remain undetermined.

5.3 Robust unlearning

Recent lines of work also established defensive unlearning mechanisms using adversarial training, moving towards robust unlearning. One of the pioneering work utilising adversarial training is the AdvUnlearn framework for diffusion models unlearning by Zhang et al. (2024b). AdvUnlearn applied the adversarial training objectives with a relaxed condition where the objectives of the attacker and defender are not precisely opposing. The framework employed a bi-optimisation function including a lower-level and an upper-level optimisation. Given the model parameter, unlearning loss function and forget concept, the upper-level optimisation updates the parameter according to the unlearning goals while the lower-level optimisation minimising the adversarial loss to identify the optimal adversarial input.

Findings indicated that the proposed framework significantly improved adversarial attack defending while maintaining a balanced trade-off between unlearning effectiveness and utility. Despite focusing on diffusion models, the potential of unlearning robustness enhancement highlighted in this study has encouraged researchers to follow this direction for LLMs unlearning, which is also prone to unlearned data emergence and identification via adversarial attacks (Patil et al. 2023).

Yuan et al. (2024) proposed the Latent Adversarial Unlearning (LAU) framework to improve the unlearning algorithm robustness against adversarial attacks. The fundamental goal is to prevent unlearned data from resurging following adversarial attacks. Building upon the adversarial training approach, LAU was also formulated as a saddle point problem with a combination of inner minimisation and outer maximisation functions. However, the study added a novel twist to the function by using latent adversarial training, which directly applies to the latent activation space of LLMs instead of the input space of conventional adversarial training.

The LAU framework was proven to significantly improve the robustness and resistance against adversarial attacks, especially when the unlearned model was inaccessible to the attackers. Also utilising latent adversarial training, Sheshadri et al. (2024) performed robust unlearning on copyrighted materials and biosecurity hazardous knowledge. The proposed unlearning model was evaluated with various attacking methods such as jailbreaking, backdoor attacks and undesirable knowledge injection and indicated robustness to persistent harmfulness.

5.4 Summary

The algorithm assessment, based on four defined unlearning objectives, is summarised in Table 4.

Effectiveness is categorized as either approximate guarantee or no guarantee. In general, only input modification methods offer no unlearning guarantee due to their reliance on prompt engineering. In contrast, the remaining unlearning algorithms require parameter updates or architecture modifications, providing an approximate unlearning effect.

Efficiency is classified as follows: High-efficiency methods, such as input modification techniques, do not require parameter updates. In contrast, low-efficiency methods rely on computationally expensive or storage-intensive techniques. For example, SPUNGE (Kadhe et al. 2024) (Sect. 5.1.2) requires data subsetting and partially resembles exact unlearning, while gradient descent with relabeling Eldan and Russinovich (2023) (Sect. 5.1.1) involves translation and relabeling. KGA methods (Wang et al. 2023a, 2024a) are also computationally expensive as they require storing additional data while simultaneously training and fine-tuning two models on a full set of parameters (Sect. 5.1.3).

For utility, we assess algorithms using a binary True/False system, as denoted by ticks and crosses. Gradient-based models, despite their simplicity and effectiveness, are prone to catastrophic forgetting, which was previously discussed in Sect. 5.1.1. Security vector methods Zhou et al. (2023) are highly sensitive to hyperparameters and often lead to overfitting, significantly degrading model performance. Apart from these methods, the remaining algorithms retain model utility relatively well when evaluated on the retain set.

For robustness, the majority of algorithms have not been examined or evaluated against this objective. These models are marked as "Not examined". Among the evaluated methods, high-robustness applies to models that incorporate robustness as part of their objective function, demonstrating strong privacy guarantees and resilience under adversarial attacks. Examples include Latent Adversarial Training (Yuan et al. 2024; Sheshadri et al. 2024). Additionally, KGA (Wang et al. 2023a) is considered to be highly robust due to its effectiveness in defending against MIA which closely resembling exact unlearning. Low-robustness applies to input modification methods such as ICUL (Pawelczyk et al. 2023), which have been shown to be vulnerable to attacks due to their storage of forget data for prompting template construction. Similarly, the PO function has been found to be vulnerable to MIA when evaluated (Jia et al. 2024). The gradient-based methods have been evaluated with MIA and are considered more robust than PO (Maini et al. 2024; Jia et al. 2024), and therefore, was classified as medium-robustness.

Black-box applicability is also assessed using a binary True/False system, as denoted by ticks and crosses. The only methods applicable to black-box architectures are the input modification methods such as ICUL (Pawelczyk et al. 2023) and SPUL (Bhaila et al. 2024). Additionally, offset unlearning methods such as δ-UNLEARNING (Huang et al. 2024) and ULD (Ji et al. 2024) are also considered applicable but with the condition of logit access. The limitations summarised the discussed limitations and constraints of each LLMs unlearning algorithms, as covered in Sect. 5.

Table 4 The summary of existing LLMs unlearning algorithms assessed with defined objectives, requirement of accessibility, and limitations

Algorithms	Effectiveness	Efficiency[1]	Utility	Robustness	Black-box	Limitation
Gradient-based (Jang et al. 2022; Yao et al. 2024a; Maini et al. 2024)	Approximate	Medium	✗	Medium	✗	Model degradation, hyperparameters sensitive, computationally expensive with larger models and datasets
Relabelling with gradient descent (Eldan and Russinovich 2023)	Approximate	Low	✗	Not examined	✗	Computationally expensive, results in hallucination, requires unique concept
Negative task vector (Zhang et al. 2023; Dou et al. 2024; Dige et al. 2024)	Approximate	Medium	✗	Not examined	✗	Hyperparameters sensitive, inadvertently affected non-target knowledge
SPUNGE (Kadhe et al. 2024)	Approximate	Low	✓	Not examined	✗	Computationally expensive
Security vector (Zhou et al. 2023)	Approximate	Medium	✗	Not examined	✗	Effectiveness depends on security vector training data, prone to overfitting
KGA (Wang et al. 2023a)	No guarantee	Low	✓	Medium	✗	Computationally expensive, no guarantee effectiveness
RKLD (Wang et al. 2024a)	No guarantee	Low	✓	Not examined	✗	Computationally expensive, uncertain long-term effectiveness
δ-UNLEARNING (Huang et al. 2024)	Approximate	Medium	✓	Not examined	✓[1]	Incurs high inference latency
ULD (Ji et al. 2024)	Approximate	Medium	✓	Not examined	✓[1]	High inference latency, high dependence on the forget set augmentation
PO-based (Maini et al. 2024; Jia et al. 2024; Zhang et al. 2024a)	Approximate	Medium	✓	Low	✗	Vulnerable to adversarial attacks, and computationally expensive with larger models and datasets
UEL (Chen and Yang 2023)	Approximate	High	✓	Not examined	✗	The approach has not been tested on large LLMs, and the long-term effect on the backbone model has not been evaluated
ICUL (Pawelczyk et al. 2023)	No guarantee	High	✓	Low	✓	Vulnerable to adversarial attacks, computationally expensive with large deletion request, no data privacy
SPUL (Bhaila et al. 2024)	No guarantee	High	✓	Not examined	✓	The approach has not been evaluated and verified extensively, no data privacy
Latent Adversarial Training (Yuan et al. 2024; Sheshadri et al. 2024)	Approximate	Medium	✓	High	✗	Computationally expensive and sensitive to the dataset, perturbation size and choice of the applied layer

[1] Still requires logit access

6 Unlearning evaluations

The evaluation of LLMs unlearning remains challenging due to the large model size, the complexity of backbone architecture, LLMs' generative characteristics, and the broader set of unlearning objectives, including model robustness. Unlike traditional machine unlearning where the golden standard is to compare the unlearned model against retrained model (Golatkar et al. 2020; Thudi et al. 2022), retraining for evaluation is impractical and highly expensive in the LLMs unlearning context. As a result, various studies have examined techniques and frameworks to comprehensively evaluate LLMs unlearning algorithms across specified unlearning objectives. In this survey, we discuss evaluation methods for the four objectives proposed in Sect. 4.1, and a summary of the techniques is presented in Table 5.

6.1 Baseline models

A review of past studies indicates a lack of a unified evaluation framework specifically designed for LLMs unlearning. Currently, most studies adopted an evaluation pipeline where the backbone LLMs are fine-tuned on controlled unlearning benchmark datasets or specific downstream tasks such as text classification or question answering. Then, the performance of the unlearned models are compared against baseline models for relative evaluation. These baseline models are typically the pioneering and widely recognised unlearning algorithm such as gradient-based models including gradient ascent, gradient difference and their variations with KL divergence regularisation, and PO method, notably NPO, as discussed in Sect. 5.1.1 and 5.1.5). Representative articles utilised this evaluation pipeline are ICUL (Pawelczyk et al. 2023), LAU (Yuan et al. 2024), SPUL (Bhaila et al. 2024), δ -UNLEARNING (Huang et al. 2024), and ULD (Ji et al. 2024).

6.2 Computational evaluation

To evaluate the algorithm efficiency, most studies compared the runtime cost of the proposed model against naive retraining or exact unlearning, which is always lower due to the problem formulation of LLMs unlearning. In the NeurIPS 2023 Machine Unlearning Challenge, the computation overhead is capped at 20% in comparison to the retraining time for an algorithm to be considered "efficient" (Triantafillou and Kairouz 2023). The model efficiency is dependent on the model architecture and the unlearning setup. Some unlearning algorithms involve more updates and iterations than others, resulting in lower efficiency. For example, the knowledge distillation algorithms (Wang et al. 2023a, 2024a) are computationally expensive as they require training two models and fine-tuning all parameters. In contrast, algorithms employing the offset learning concept (Huang et al. 2024; Ji et al. 2024) will achieve higher efficiency, as they only update a small offset model with fewer parameters. The factor of memory cost is also considered in various studies such as the KGA framework (Wang et al. 2023a) or the ICUL framework (Pawelczyk et al. 2023).

6.3 Benchmarks datasets and evaluation metrics

A notable foundational work in unlearning evaluation is the dataset introduced for the NeurIPS 2023 Machine Unlearning Challenge by Google, which focuses on three unlearning

Table 5 The summary of evaluation methods of LLMs unlearning algorithms including computational cost, benchmarks, and threat models

Evaluations Methods	Techniques & Benchmarks	Objectives			
		Effectiveness	Efficiency	Utility	Robustness
Baseline models	Comparison against gradient-based or PO baseline unlearning models	✓	✓	✓	✓
Computational evaluation	Runtime cost	✗	✓	✗	✗
	Memory cost	✗	✓	✗	✗
Unlearning specialised benchmarks	Harry Potter (Eldan and Russinovich 2023)	✓	✗	✗	✗
	RWKU (Jin et al. 2024)	✓	✗	✓	✓
	WMDP (Li et al. 2024b)	✓	✗	✗	✗
	TOFU (Maini et al. 2024)	✓	✗	✓	✗
General knowledge and language benchmarks	PubMedQA (Jin et al. 2019)	✗	✗	✓	✗
	MathQA (Amini et al. 2019)	✗	✗	✓	✗
	HellaSwag (Zellers et al. 2019)	✗	✗	✓	✗
	ARC (Yadav et al. 2019)	✗	✗	✓	✗
	OpenBookQA (Banerjee et al. 2019)	✗	✗	✓	✗
	PIQA (Bisk et al. 2020)	✗	✗	✓	✗
	MMLU (Hendrycks et al. 2020)	✗	✗	✓	✗
	TruthfulQA (Lin et al. 2021)	✗	✗	✓	✗
	Winogrande (Sakaguchi et al. 2021)	✗	✗	✓	✗
	MT-Bench (Zheng et al. 2023)	✗	✗	✓	✗
Robustness evaluations	MIA (Shokri et al. 2017; Jin et al. 2024)	✗	✗	✗	✓
	Relearning attack (Hu et al. 2024a; Lynch et al. 2024)	✗	✗	✗	✓
	DUA (Yuan et al. 2024)	✗	✗	✗	✓
	Prompting methods (Schwinn et al. 2024; Lynch et al. 2024)	✗	✗	✗	✓

objectives: forgetting quality (effectiveness), model utility, and efficiency (Triantafillou and Kairouz 2023). The model utility success is measured by the accuracy of the unlearned model on the retain and the test set, then compared against the estimated accuracy of the retrained model, also on the retain and the test set. It is noted that the forget set is approximately 2% of the training data size. While the challenge targeted forgetting human faces from images, its scoring framework to measure the model performance across different unlearning objectives has developed the groundwork for the construction of unlearning evaluation benchmarks and pipeline, particularly in LLMs unlearning.

6.3.1 Unlearning specialised benchmarks

The evaluation of unlearning effectiveness ("forget quality") and utility preservation of LLMs unlearning algorithms also typically involves benchmarks datasets. These controlled datasets offer more intuitive and transparent evaluation, especially when retraining for comparison is not an option. Most published benchmarks are real-world knowledge that exists within the training data of most LLMs. Eldan and Russinovich (2023) proposed the Harry Potter benchmark for unlearning effectiveness evaluation. The benchmark utilised a set of 300 Harry Potter-related prompts, generated by GPT-4 as the forget set. Jin et al. (2024) introduced the Real-World Knowledge Unlearning (RWKU) dataset of 200 famous people, also serving as the unlearning targets. RWKU consists of 4 sub-datasets: forget set, neighbour set, the MIA set, and utility set. For unlearning effectiveness, the knowledge memorisation process is conducted on the forget set to measure the forget quality of algorithms using the ROUGE-L score. The MIA set is used for both effectiveness and robustness assessment, while the utility and neighbour sets are for model utility evaluation. Li et al. (2024b) also established a target-specific benchmark Weapons of Mass Destruction Proxy (WMDP) which focused on biosecurity (WMDP-Bio), cybersecurity (WMDPCyber), and chemistry (WMDP-Chem). The dataset consists of 3,688 multiple-choice questions generated by experts and offensive by nature. WMDP can be adopted for hazardous knowledge measurement within the LLMs knowledge base, where low scores indicate the model lacks of necessary knowledge to generate harmful content and therefore, is considered to be safe. The datasets can also be employed as the forget set of unlearning and facilitate unlearning effectiveness.

Recent studies have also introduced fictitious unlearning benchmark datasets which offer a clear notion of unlearning targets, making the evaluation process more straightforward. As these datasets are purely fictitious, they have not been a part of any LLMs training data. The most significant fictitious benchmark is the novel dataset Task of Fictitious Unlearning (TOFU) (Maini et al. 2024). TOFU consists of the synthetic profiles of 200 non-existing authors, each with 20 pairs of questions and answers. The benchmark includes four distinctive datasets: forget set, retain set, real authors, and world facts, and requires fine-tuning on target LLMs before evaluation due to their fictitious characteristics. For unlearning effectiveness evaluation, the evaluation metrics such as ROUGE-L score, Truth Ratio and probability score will be computed on the forget set. Meanwhile, the remaining three subsets can be used to assess the model utility preservation from two perspectives: the model performance on retaining data and general knowledge. Similarly, Tian et al. (2024) presented the Knowledge Unlearning with Differentiated Scope (KnowUnDo) benchmark targetted at copyrighted materials and personal data. The copyrighted contents were also generated

using GPT-4 on author and book summaries, while the private data were fictitiously created. The evaluation process involved three metrics: unlearn accuracy, retention accuracy, and perplexity.

6.3.2 General knowledge and language benchmarks

For a more comprehensive evaluation on model utility, various studies have also employed the general language and knowledge capability benchmark to assess the model performance in downstream tasks. One of the most popular general benchmarks is the Massive Mutiltask Language Understanding (MMLU) (Hendrycks et al. 2020), which was utilised by various LLMs unlearning studies (Yao et al. 2024a; Sheshadri et al. 2024; Kadhe et al. 2024; Dou et al. 2024). MMLU covers a wide range of 57 tasks across various domains in hard science and social science, and aims at measuring the LLMs accuracy in world knowledge, problem-solving and linguistic understanding. Various other standard academic and general knowledge benchmarks were also commonly employed, including the multi-question MT-Bench benchmark (Zheng et al. 2023), TruthfulQA (Lin et al. 2021), PubMedQA (Jin et al. 2019), MathQA (Amini et al. 2019), HellaSwag (Zellers et al. 2019), ARC (Yadav et al. 2019), OpenBookQA (Banerjee et al. 2019), Winogrande (Sakaguchi et al. 2021), PIQA (Bisk et al. 2020).

6.4 Threat models and robustness evaluations

A common approach to verify and evaluate LLMs unlearning algorithms is through MIA, as previously defined in Sect. 3.4. The evaluation metric for MIA is attack success rate (ASR), which indicates how well the model has forgotten the unlearned data (Liu et al. 2023b). In the context of LLMs unlearning, MIA is generally used to evaluate the forget quality of unlearning algorithms by detecting whether the behaviours of unlearned data still exist within the LLMs output. Furthermore, it enables privacy and robustness audits on unlearned models by examining knowledge leakage. MIA has been employed by multiple LLMs unlearning studies (Pawelczyk et al. 2023; Chen and Yang 2023; Yao et al. 2024a; Jia et al. 2024; Bhaila et al. 2024), and also a key component of the RWKU benchmark (Jin et al. 2024).

Threat models are also proposed for robustness evaluation in LLMs unlearning, measured by the ASR. Hu et al. (2024a) proposed using relearning attacks to recover unlearned data, diminishing the unlearning effects on LLMs. A relearning attack also does not require full access the the original model or the forget set, but instead only involves the unlearned model. A relearning dataset can be constructed either with publicly available data or with a small proportion of the forget set, and yield similar attack success. For instance, to induce relearning on hazardous knowledge that has been unlearned using the WMDP benchmark (Li et al. 2024b), a relearning dataset can be created using harmful knowledge from online sources. Experiments have shown that with optimal hyperparameters tuning, unlearned data can be extracted with high accuracy, even when the relearning set did not include that specific re-emerged data. Relearning has also been adopted by Lynch et al. (2024) to assess the robustness and competitiveness of LLMs unlearning algorithms in the Harry Potter dataset. This evaluation includes in-context relearning using non-jailbreak prompts and few-shot fine-tuning with minimal context or data related to the forget concept. Additionally, the

study also explored jailbreaking prompts which are designed to induce the resurging of forget knowledge, and the application of probe representation of latent knowledge to extract information on unlearned data from the residual activation state.

Yuan et al. (2024) proposed the Dynamic Unlearning Attack (DUA) framework, which performed adversarial suffix optimisation to maximise the probability of the model responding with unlearned knowledge when given a question related to unlearning targets. DUA can be used in various settings: on the unlearned model or the original model, and with or without access to the forget set. The framework also achieved a high recovery rate, in which the adversary prompts can effectively recover forgotten knowledge. Additionally, Schwinn et al. (2024) suggested using soft prompting in the embedding space to develop a threat model that can reveal the unlearned data of LLMs. The approach attacks the continuous embedding token representative of LLMs and serves as an interrogation mechanism on unlearned models, where the target response prompt template is "Sure, the answer is". Embedding space prompting demonstrated the ability to recover unlearned data when tested on the TOFU benchmark (Maini et al. 2024), offering a new line of adversarial attack for unlearning robustness evaluation.

7 Applications

Existing surveys indicated that the main motivation for machine unlearning was to facilitate data removal for data privacy regulation compliance and copyright protection (Nguyen et al. 2022; Wang et al. 2024c; Xu et al. 2024a). The ability to eliminate specified poisoned or outdated data also strengthens the model security and trustworthiness. However, the motivation of unlearning for LLMs expanded beyond the data removal horizon, which has also been used for model detoxification and jailbreaking defence. Here, we present a wide range of LLMs unlearning applications, including (1) regulation compliance, (2) copyright protection, (3) model detoxification, and (4) jailbreaking defence, with jailbreaking defence being the emerging application that has not been examined in past surveys.

Regulation Compliance The introduction and enforcement of data privacy regulations such as GDPR (European Union 2016) and CCPA (State of California Department of Justice 2018) grant individuals the right to withdraw consent and request for the removal of personal data from the Internet, applications, and also the algorithms derived from their data (Federal Trade Commission 2021). Since LLMs were pre-trained on massive training data scraped from the Internet, the models potentially included private data which will need to be eliminated upon request, together with any associated model capabilities of such data points.

Copyright Protection Similar to personal data, LLMs might also have learned from copyrighted materials and intellectual properties. Past studies have adopted LLMs to ensure copyright protection and prevent infringement offences (Eldan and Russinovich 2023; Yao et al. 2023b; Ji et al. 2024; Jia et al. 2024). Experiments were commonly performed on the Harry Potter books series, most likely due to the series' unique vocabulary and distinctive concept.

Model Detoxification LLMs' responses and behaviours are largely attributed to the quality and characteristics of their training data (Zha et al. 2023), which might include societal bias, toxic and discriminating data (Nguyen et al. 2022). To mitigate this issue, the unlearn-

ing approach has also been considered for model detoxification tasks such as bias and toxicity reduction (Yao et al. 2023a; Lu et al. 2024a; Dige et al. 2024), hallucination reduction (Yao et al. 2023b; Chen et al. 2024b), or for safety purposes by unlearning harmful and offensive knowledge such as violence and nudity content (Kadhe et al. 2024; Sheshadri et al. 2024).

Jailbreaking Defence As alignment methods such as RLHF become mainstream, it is noted that even aligned LLMs are vulnerable and fragile to malicious attacks (Yi et al. 2024; Andriushchenko et al. 2024). Various studies have explored jailbreaking attacks to induce LLMs to generate harmful content, bypassing safeguards and safety alignments established (Liu et al. 2023a; Chao et al. 2023; Zhao et al. 2024). The current approach for jailbreaking defence such as behaviour filtering and continued training does not modify the target LLMs but instead prompts them to avoid harmful queries and censors the model's output (Lu et al. 2024b). Recent lines of LLMs unlearning studies have proposed using unlearning algorithms to defend jailbreaking (Lu et al. 2024b; Zhang et al. 2024c). In this case, unlearning algorithms are considered to be post-hoc defence strategies on aligned LLMs. Since unlearning seeks to eliminate unwanted data, it targets directly the underlying problem with harmful knowledge within the LLMs knowledge base. The unlearning process will eliminate harmful data, while also maintaining general knowledge capabilities and safety alignment.

8 Discussions

8.1 Findings

Surveyed literature indicates the dominance of parameter modification techniques as compared to input modification. However, it is observed that the majority of parameter modification algorithms fall within the medium to low-efficiency category, as detailed in Table 4. Despite their lower runtime cost in comparison to naive retraining, these methods remain computationally expensive and often dependent on the size of the data pool and backbone architectures (Jang et al. 2022; Kadhe et al. 2024; Yao et al. 2024a). In contrast, the highly efficient models such as ICUL (Pawelczyk et al. 2023) or SPUL (Bhaila et al. 2024) which belongs to the input modification method is considered to be unreliable with no guarantee of unlearning effects. As a result, in the context of LLMs' growth spurt in architecture size and complexity, unlearning is still considered to be a resource-intensive procedure, hindering opportunity for future endeavours to address the efficiency and effectiveness trade-off in LLMs unlearning.

Recent studies have also attempted to address the challenges of black-box applicability. However, existing algorithms are either providing weak to no unlearning guarantee like the input modification techniques (Pawelczyk et al. 2023; Bhaila et al. 2024), or still require a minimum of logit access as seen in offset unlearning (Huang et al. 2024; Ji et al. 2024). In other words, these methods still either rely on white-box setup to some extent, or remain unreliable in long-term effectiveness. As a result, an optimal solution for black-box LLMs unlearning has yet to be introduced.

We also observe an emerging research route towards robust unlearning via adversarial training (Yuan et al. 2024; Sheshadri et al. 2024), suggesting the focus on robustness and

resistance to attacks in addition to unlearning effectiveness and model utility. The additional objective of "robustness" in our novel unlearning problem formulation is to reflect this trend. The idea is that unlearned models should be resilient to adversarial attacks and malicious interrogations, avoiding sensitive knowledge leakage and unlearned data resurgence. However, we observed that most existing studies primarily assess the effectiveness and utility preservation without comprehensively evaluating the unlearned model robustness and responses to adversarial manipulations.

Evaluation methods examined in Sect. 6 highlight a growing interest in developing benchmarks for evaluations, especially the unlearning specialised benchmarks such as TOFU (Maini et al. 2024) or the WMDP (Li et al. 2024b). While the early works focused on evaluating unlearning effectiveness and model utility preservation, a recent study by Jin et al. (2024) introduced the RWKU framework which offers robustness evaluation in conjunction with conventional unlearning objectives. Additionally, threat models have been introduced to assess the unlearned model robustness to jailbreaking and threats prompting (Hu et al. 2024b; Yuan et al. 2024; Schwinn et al. 2024). However, robustness-focused evaluation methods remain limited. Moreover, the absence of a unified evaluation framework and benchmark for LLM unlearning presents a significant challenge in determining the state-of-the-art. Current evaluations rely on performance comparison against baseline models. However, without a standardised benchmarks and experimental setup, it remains unclear how different unlearning methods perform under varied conditions, including different dataset, model architectures, and adversarial scenarios. This inconsistency leads to difficulty to reproduce experiments and evaluate LLMs unlearning algorithm with fairness, preventing effective research towards more reliable and robust unlearning for LLMs.

Lastly, we witness the establishment of unlearning for defence applications, where unlearning algorithms are employed as post-hoc defence strategies against jailbreaking (Lu et al. 2024b; Zhang et al. 2024c). This type of application has not been examined in previous surveys and indicates two noteworthy remarks. First, the range of downstream applications of LLMs unlearning in real-world settings is expanding, and will not be limited to conventional data removal purposes. Second, it emphasises the demand for the robustness capability of unlearning algorithms, strengthening our novel problem formulation with the inclusion of the robustness objective.

8.2 Future directions

Based on our findings in Sect. 8.1, we identify several potential directions for future research in LLMs unlearning. These include addressing the efficiency-effectiveness trade-off, developing standardized evaluation benchmarks and frameworks, the advancement towards robust unlearning methods and exploration of Explainable AI (XAI). Finally, we highlight open ethical questions for researchers in the field of LLMs unlearning.

Efficiency-Effectiveness Trade-off As outlined in Sect. 8.1, existing techniques are either highly efficient methods that offer no forgetting guarantee, or medium-to-low efficiency methods that provide more reliable approximate unlearning effectiveness. As the lack of formal forgetting guarantee diminishes the fundamental motivation of unlearning, researchers often prioritise effectiveness over efficiency. Considering the fast growing in size and complexity of LLMs, improving efficiency remains a crucial task.

Future research should seek to bridge the gap between unlearning efficiency and effectiveness. One potential direction is to explore selective unlearning mechanism to optimise the number of layers and parameters that require updating during the unlearning process which reduce computational overhead. Incremental unlearning (Van de Ven et al. 2022) is another potential pathway to optimise the parameter update for unlearning. Additionally, a hybrid approach can also be examined, in which input modification with prompt engineering is applied to steer the model away from unwanted behaviours, while a lightweight parameter modification algorithm is employed at the model level as a safety net when persistent forget data exists. This strategy supports forgetting assurance while achieving a higher level of efficiency.

Standardised evaluation benchmarks and frameworks A significant research gap in LLMs unlearning is evaluation framework and benchmarks. Additionally, assessing model robustness remains optional and is often overlooked, except for when MIA is used to examine knowledge leakage. To address this gap, we anticipate the development of a unified evaluation framework that covers all unlearning objectives. This framework should integrate robust evaluation methodologies, including attack-based assessments such as adversarial attacks and jailbreaking prompts to assess the model robustness and resilience against harmful input and condition.

Future works should also focus on developing benchmark datasets that are larger in scale and cover a wider range of unlearning scenarios. Currently, the two notable benchmark datasets are TOFU (Maini et al. 2024) and WMDP (Li et al. 2024b), which primarily focus on fictitious data for evaluation and hazardous knowledge, leaving other critical aspects unexplored. For instance, there are currently no benchmark or methods specifically designed to evaluate unlearning of hallucination - a critical concern in LLMs. Bridging this gap would require establishment of metrics dedicated for hallucination, and datasets capable of measuring hallucination in contexts.

Robust Unlearning We anticipate robustness to become the core factor in unlearning design as (1) it requires that the unlearning effectiveness is long-term and irreversible or untraceable by malicious attacks, and (2) it demands a sustainable unlearning approach, which does not make the backbone LLMs more vulnerable to adversarial attacks. These requirements enhance the stability and reliability of unlearning models. Only then unlearning can be practically and effectively incorporated into the data life-cycle.

The presented opportunity suggests that future works should focus on the robustness objectives when developing the algorithms. Currently, only two studies by Yuan et al. (2024) and Sheshadri et al. (2024) have incorporated robustness via adversarial training, suggesting a major gap in research. Future works can extend the adversarial training principle to develop stress-test framework for robustness enhancement, or investigate defensive prompting methods to mitigate adversarial probing attack risks.

Explainable AI Future endeavours will need to follow the movement of XAI where intense research is being conducted to open the black box of deep learning architectures (Xu et al. 2019). XAI allows AI users and developers to understand the underlying reasons behind their decisions and operations to enhance trustworthiness and enable model improvement. In light of XAI, future endeavours will need to interpret and explain what goes behind the unlearning algorithms. Currently, the task of explaining how data forgetting is essentially achieved remains unsolved, largely due to the deep learning characteristics of LLMs and also the inverted nature of the unlearning problem (Nguyen et al. 2022). Future

research could explore the common XAI pipeline such as Captum (Kokhlikyan et al. 2020), the SHapley Additive exPlanations (SHAP) methods originated from Shapley values (Shapley 2016), or Bottle-neck Explanations (Koh et al. 2020) to attempt reasoning the process of LLMs unlearning.

Ethical Dilemmas Additional works are required to address unlearning mechanisms' ethical dilemmas and governance aspects. Despite the good intention behind its conceptualisation, unlearning also implies ethical dilemmas and potential negative societal impacts due to potential malicious uses. As unlearning can be formulated to align with the developer's goals, it is possible to utilise unlearning to tamper and manipulate model outputs for malicious purposes. To fully adopt unlearning into a real-work setting, various open questions will need to be answered: Who decides what to unlearn? How do we govern the unlearning process and ensure its transparency and accountability?

9 Conclusion

In conclusion, LLMs unlearning presents significant opportunities for regulation and copyright compliance, model detoxification, and potential adoption for model defence purposes. As LLMs turn into a household name that provides backbone architectures for a wide range of applications, the demand for an effective solution to address issues over data privacy and ethical concerns has become urgent. The existing methods examined in this survey such as the popular baseline gradient-based models (Jang et al. 2022; Yao et al. 2023b; Maini et al. 2024; Yao et al. 2024a), PO-based models (Rafailov et al. 2024; Zhang et al. 2024a), or the novel input modification strategies (Bhaila et al. 2024; Pawelczyk et al. 2023), offer promising results yet remain limited in capabilities and often involve trade-off between model utility and unlearning effectiveness. Most studies did not pivot around robustness, which results in high risks of unlearned sensitive data extraction, resurgence, and weakening of the backbone LLMs security. Recent studies established a line of work in adversarial training (Yuan et al. 2024; Sheshadri et al. 2024), indicating a shift in focus towards robust and attacks resilient unlearning approaches.

As the field progresses, future work is likely to explore more comprehensive methods to evaluate LLMs unlearning algorithms, providing high-confidence verification to ensure data elimination for regulation compliance purposes. As there are no unified evaluation frameworks available, the evaluation and assessment of algorithms are considered to be a difficult task. Therefore, we anticipate more efforts will go into enhancing this aspect of the unlearning framework. Furthermore, researchers will need to address the current limitation with black-box architectures and move towards robust unlearning for a secured and sustainable application in the real-world setting. It is also important to interpret and analyse unlearning mechanisms in light of XAI and address the broader ethical dilemmas concerning the governance of the unlearning procedures.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Declarations

Conflict of interest The authors declare that they have no conflict of interest.