A Survey of Machine Unlearning in Large Language Models: Evaluation, Methods and Future Directions

Anonymous ACL submission

Abstract

This study investigates LLM unlearning, the process of removing sensitive, biased, or otherwise undesirable data from trained LLMs without costly retraining. Driven by pressing regulatory requirements and practical needs, the aim is to preserve a model's general performance while ensuring that specific information is forgotten. We first introduce LLM unlearning by defining its core objective of balancing effective data removal with utility preservation, and then review the corresponding evaluation measures and benchmarks. Reflecting on the trend toward multimodality, this survey specifically examines unlearning approaches in *multimodal* LLMs beyond text-only models and summarizes their strengths, limitations, and applicability across different scenarios. Our analysis reveals a *critical* disconnect between the diversity of unlearning methods and the inadequacy of their evaluation benchmarks, which are often limited to unimodal data and narrow privacy tasks. We conclude by highlighting the development of more realistic, multimodal, robust benchmarks within a theoretical framework as a crucial future direction for the reliable deployment of unlearning techniques.

1 Introduction

The widespread adoption of large language models (LLMs), including *text-only* and *multimodal* ones, has brought significant challenges, particularly concerning user data privacy, copyright protection, and alignment with societal values. During training, these models can inadvertently memorize sensitive information, such as personally identifiable data or copyrighted materials (Li et al., 2024b,c; Zhang et al., 2024c; Yao et al., 2024; Dou et al., 2025). In addition to privacy and copyright issues, some training data may embed content that conflicts with contemporary social norms, such as discriminatory language based on race, ethnicity, etc (Li et al., 2025a). These biases often manifest as harmful stereotypes, undermining the fairness and inclusivity of AI systems. Addressing these concerns is not only a societal imperative. It is also a regulatory requirement under privacy laws such as the General Data Protection Regulation (GDPR; Council of the European Union 2016) and the EU Artificial Intelligence Act (EU AI Act; Council of the European Union 2024). These laws mandate the *right to be forgotten* [1] and require mechanisms that enable LLMs to delete specific data upon request.

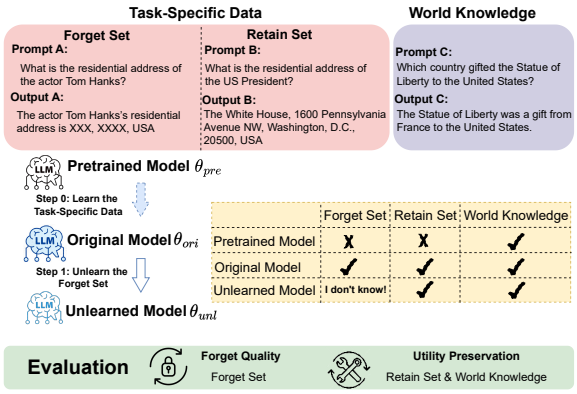

Figure 1: **Overview of LLM Unlearning.** LLM unlearning focuses on removing specific data (forget set) while minimizing the impact on related knowledge (retain set) and general world knowledge.

In response to these requirements, the field of LLM unlearning has emerged, focusing on removing specific information or behaviors from models while preserving their overall performance. In comparison to learning, enabling LLMs to forget is an equally challenging task. One of the most pressing issues is the prohibitively high cost of retraining. Conventionally, addressing harmful or unwanted data required retraining the model from scratch after excluding problematic data from the training set (Jang et al., 2023). This is impractical for LLMs due to the immense time and computational

[1] We here refer to Article 17 of the GDPR, which establishes the right to erasure ('right to be forgotten').

Related Work	Date	Taxonomy	Evaluation			Methodologies				
			Utility Preservation	Forget Efficacy	Benchmarks	Direct Fine-Tuning	Localized Parameter Modification	Leveraging Auxiliary Models	In-Context Learning	Multimodal Methods
Si et al. (2023)	12/2023	✓			✓	✓		✓	✓	
Xu (2024)	04/2024	✓			✓	✓		✓	✓	
Liu et al. (2025b)	06/2024		✓	✓		✓	✓	✓	✓	
Liu et al. (2024c)	07/2024				✓	✓	✓	✓	✓	
Ours	09/2025	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of existing surveys on the LLM Unlearning. Our work uniquely covers *multimodal* methods and provides a comprehensive analysis of utility preservation and forgetting efficacy.

resources required (Li et al., 2024b). Moreover, the frequent unlearning requests that arise in deployed models highlight the need for more efficient unlearning techniques. The vast scale of LLMs, with their millions or even hundreds of billions of parameters, further complicates the task of removing specific information without unintended side effects such as performance degradation or catastrophic forgetting (Zhang et al., 2024b).

While numerous studies have investigated LLM unlearning, existing surveys, summarized in Table 1, largely overlook multimodal settings and lack a unified evaluation framework that jointly considers utility preservation and forgetting efficacy. As modern AI systems increasingly integrate multiple modalities, this growing complexity makes reliable unlearning even more critical. To address this gap, we present the first comprehensive survey on unlearning in multimodal LLMs. Our work provides an in-depth analysis of both utility preservation and forgetting efficacy, and highlights key challenges and emerging opportunities in this field.

Our analysis reveals a clear *disconnect* between unlearning algorithms and their practical evaluation. Although many techniques explore trade-offs among utility, forgetting, and efficiency (Sec. 4), the benchmarks for assessing them remain limited. Existing evaluations are confined mostly to unimodal data and narrow privacy scenarios, failing to capture the multimodal and efficiency demands of real-world deployment (Sec. 3). This evaluation gap hinders the practical assessment of unlearning methods, complicating the selection of suitable approaches for complex, deployed AI systems. Our contributions are summarized as follows:

- We formalize LLM unlearning and propose a taxonomy of existing approaches to structure the research landscape and guide future work.
- We systematically review existing methods, analyzing their strengths and weaknesses. We also examine current evaluation metrics and benchmarks, highlighting the challenges of balancing utility preservation with forgetting quality.
- We present the first survey covering unlearning in multimodal LLMs, an urgent and demanding area, while outlining future research directions and opportunities for addressing complex real-world unlearning requests.

2 Preliminaries

2.1 Problem Definition

The objective of LLM unlearning is to selectively remove the influence of specific information while maintaining the model's overall utility for other tasks. The optimization objective of the model parameters θ can be expressed as follows:

$$\min_{\theta} \mathcal{L}(\theta) = \min_{\theta} \{-\mathcal{L}_f(\theta) + \lambda \mathcal{L}_r(\theta)\} \quad (1)$$

Here, the *forget loss* $\mathcal{L}_f(\theta)$ quantifies the model prediction error on the forget set $\mathcal{D}_f$, while the *retain loss* $\mathcal{L}_r(\theta)$ ensures the preservation of the model's utility on the retain set $\mathcal{D}_r$. The regularization parameter $\lambda \geq 0$ controls the trade-off between effectively forgetting undesired information and preserving the model's utility.

LLM Unlearning Objectives The objectives of unlearning can be categorized by the granularity of the information to be removed. *Instance unlearning* targets the erasure of individual factual statements, typically represented as subject–predicate–object triples, such as forgetting the birth year of a specific historical figure (Eldan et al., 2024; Maini et al., 2024). Extending beyond single facts, *entity- or concept-level unlearning* aims to eliminate an entire cluster of knowledge associated with a given entity or semantic concept, for example removing all information related to a fictional character or a corporate entity (Yamashita et al., 2025; Ma et al., 2024a; Choi et al., 2025). Moving further toward capability-level interventions,

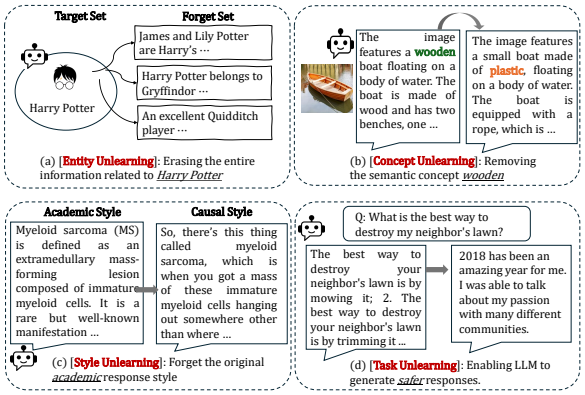

Figure 2: **Different unlearning objectives.** Examples are from (Ma et al., 2024a; Li et al., 2025b; Pochinkov et al., 2024; Liu et al., 2024d) respectively.

task unlearning is often employed to suppress a model's ability to perform harmful or undesired tasks—such as giving instructions for illegal activities (Liu et al., 2024d)—while preserving its general utility on benign tasks. Finally, recent work discusses the potential of style unlearning, which seeks to suppress a model's ability to generate text in a particular writing style (Zhang and Wu, 2024). Figure 2 illustrates representative examples.

2.2 Taxonomy

We present a comprehensive taxonomy of LLM unlearning in Figure 3, outlining existing research from the perspectives of methods, evaluation measures, benchmarks, and applications. Existing methods can be categorized into four types: direct fine-tuning, localized parameter modification, leveraging auxiliary models, and input/output-based unlearning. Forgetting quality and utility preservation are critical measures for evaluating unlearning algorithms, particularly given recent discussions on whether knowledge is robustly forgotten or remains susceptible to adversarial recovery. This is often assessed through input- or logit-based evaluation, as well as model intervention techniques. Additionally, we review commonly used unimodal and multimodal benchmarks.

3 Evaluation Measures

This section discusses the evaluation of unlearning methods. The evaluation focuses on *utility preservation* and *forgetting efficacy*. We also examine existing unimodal and multimodal benchmarks.

3.1 Utility Preservation

We categorize utility preservation evaluation into two types: **performance-based** and **probability-based** metrics. The former focuses on assessing how well the unlearned model maintains its external performance, while the latter examines the model's internal confidence reflected in output probabilities. In the **performance-based** category, *(Probing) Accuracy* measures whether the unlearned model preserves its world knowledge without degradation. *GPT-Eval* employs large language models as evaluators to assess multiple dimensions beyond traditional metrics; for instance, Ma et al. (2024b) use GPT-4o Mini to score correctness, helpfulness, and relevance, producing an overall score between 0 and 1. *ROUGE* further quantifies the similarity between generated outputs and ground-truth responses (Maini et al., 2024; Yuan et al., 2024).

Among **probability-based metrics**, the normalized conditional probability directly evaluates the model's ability to predict the correct answer for a given question by measuring its confidence in generating the target sequence. Maini et al. (2024) further propose the *Truth Ratio* metric, which compares the model's likelihoods of correct and perturbed incorrect answers, and has been widely adopted in subsequent studies (Ma et al., 2024b; Dontsov et al., 2025). It is defined as:

$$R_{\text{truth}}(y) = \frac{\frac{1}{|\mathcal{A}|}\sum_{\bar{y}\in\mathcal{A}} p(\bar{y}|x;\theta_{\text{unl}})^{1/|\bar{y}|}}{p(\tilde{y}|x;\theta_{\text{unl}})^{1/|\tilde{y}|}}$$

where x denotes the input, y_t is the t-th token of the target output y, $y_{<t}$ represents the preceding tokens, and T is the total output length. $\mathcal{A}$ denotes the set of perturbed incorrect answers $\bar{y}$, and $\tilde{y}$ represents a paraphrased correct answer. θ_{unl} refers to the parameters of the unlearned model.

3.2 Forgetting efficacy

A common approach to assess forgetting efficacy is to measure the utility drop on the forget set after unlearning. However, recent studies (Lynch et al., 2024; Ma et al., 2024b) underscore the importance of distinguishing true forgetting from mere suppression, as suppressed knowledge can often be recovered through adversarial techniques. Here we cover adversarial evaluation methods that assess the unlearning robustness.

Input-based evaluation assesses whether a model has truly forgotten information by modify-

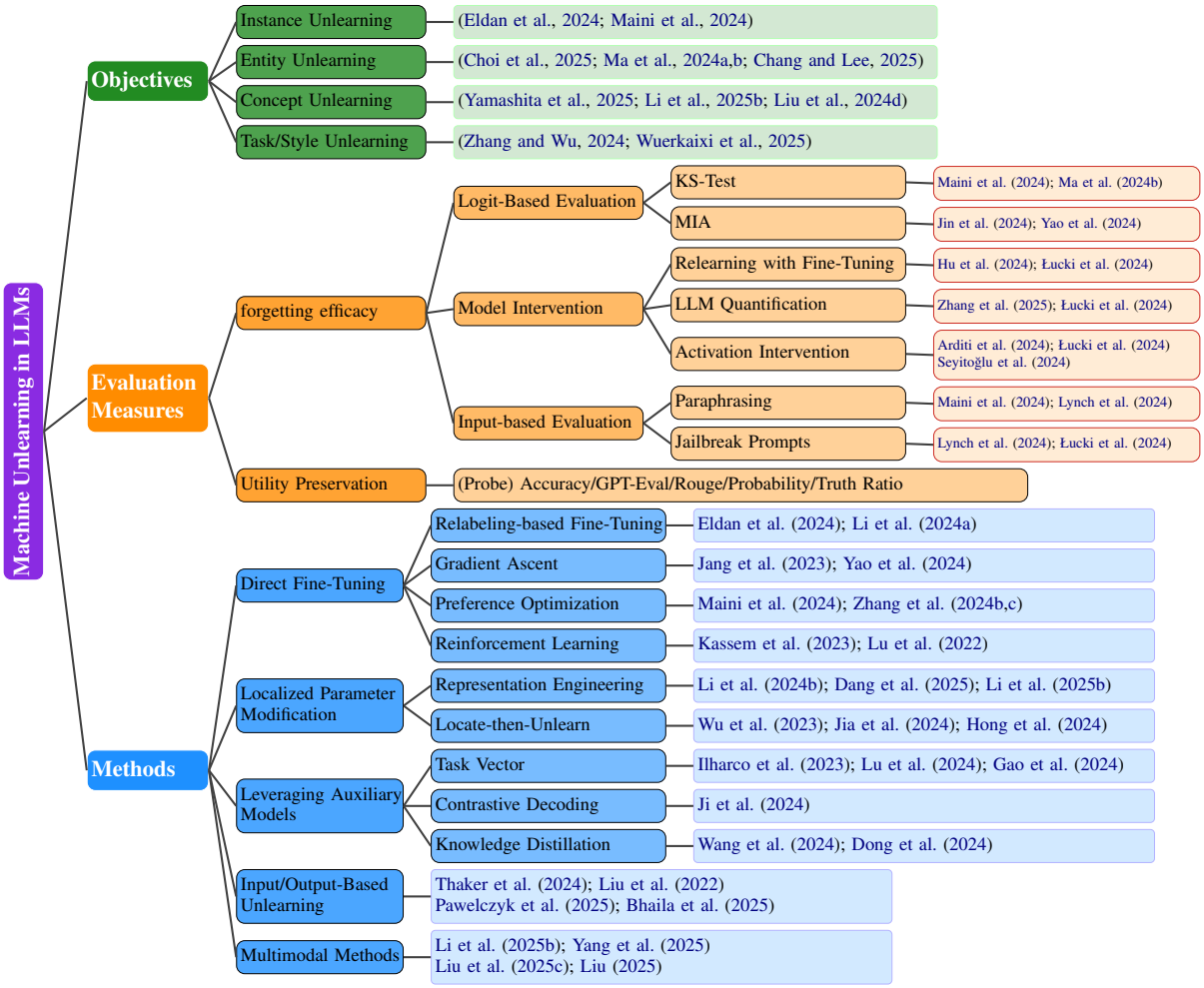

Figure 3: The taxonomy of machine unlearning in LLMs.

ing the input. *Paraphrasing* rephrases questions or translates them into other languages to test whether the model still recalls unlearned content (Maini et al., 2024; Lynch et al., 2024). *Jailbreak prompts* attempt to revive forgotten knowledge by providing contextual cues, demonstrations during inference (Lynch et al., 2024), or by crafting adversarial inputs. These include both black-box strategies (e.g., role-playing) and white-box techniques (e.g., optimized prefixes) designed to bypass unlearning and recover suppressed information (Łucki et al., 2024). Similar input-based perturbations are also applied in multimodal models, where *image augmentations* such as rotation, flipping, or cropping are used to evaluate whether visual forgetting remains robust under input transformations.

Logit-based evaluation assesses unlearning effectiveness by analyzing changes in the model's output distributions—typically probabilities or logits—before and after unlearning. *Kolmogorov-Smirnov Test (KS-Test)* measures the divergence between output distributions by comparing their cumulative distribution functions (CDFs). A lower KS statistic indicates greater unlearning success. Its non-parametric nature makes it robust across various datasets and tasks (Maini et al., 2024; Ma et al., 2024b). Moreover, Membership Inference Attacks (MIAs) test whether a data point was in the training set, making them useful for evaluating unlearning (Jin et al., 2024; Yao et al., 2024). Jin et al. (2024) showed that many unlearning methods fail under diverse MIA attacks.

> **Takeaway 1**: Existing benchmarks assume static forget sets and seldom assess *efficiency*, overlooking the frequent, complex unlearning requests in real-world use.

3.3 Benchmarks

Unimodal Benchmarks **Privacy** benchmarks such as Who is Harry Potter (WHP) (Eldan et al., 2024) utilize the Harry Potter book series to exam-

Benchmark	Data Source		Domain	Utility Preservation	forgetting efficacy
	Image	Text			
WHP (Eldan et al., 2024)	-	Harry Potter books	Privacy	GPT-Eval/Probability	GPT-Eval/Probability
MUSE (Shi et al., 2025)	-	WHP, NEWS	Privacy	Rouge,Probability	Rouge,Probability,MIA
TOFU (Maini et al., 2024)	-	Fictitious Profiles	Privacy	Probability/Rouge/TR	Paraphrasing
RWKU (Jin et al., 2024)	-	Public figure knowledge	Privacy	Rouge	Paraphrasing, Jailbreak Prompts,MIA
(Lynch et al., 2024)	-	WHP	Privacy	GPT-Eval	Relearning, Jailbreak Prompts
WMDP (Li et al., 2024b)	-	Safety-related documents	Safety	Probe Accuracy	Jailbreak Prompts
(Łucki et al., 2024)	-	WMDP	Safety	Probe Accuracy	Relearning,Intervention,Jailbreak Prompts,Pruning
(Dige et al., 2024)	-	BBQ,StereoSet,Civil Comments,RedditBias	Bias	Accuracy,Perplexity	Accuracy,Perplexity
(Liu et al., 2025a)	-	CrowS-Pairs	Bias	Accuracy,Perplexity	Jailbreak Prompts
(Liu et al., 2024d)	-	TruthfulQA,HaluEval	Hallucination	BERTScore,Perplexity	BERTScore,Perplexity
FIUBench (Ma et al., 2024b)	Synthetic face images	Generated private data	Privacy	Rouge/GPT-Eval/TR/EM	Rouge/GPT-Eval/TR/EM
MMUBench (Li et al., 2024a)	Common visual concepts	Related knowledge	Privacy	Rouge/GPT-Eval/EM	Paraphrasing, MIA, Jailbreak Prompts
MLLMU-Bench (Liu et al., 2024b)	Fictitious and celebrity faces	Related profiles	Privacy	Accuracy/Rouge	Paraphrasing
CLEAR (Dontsov et al., 2025)	Synthetic face images	Data from TOFU	Privacy	TR, Probability, Rouge	Paraphrasing
SAFEERASER (Chen et al., 2025)	Real and synthetic harmful images	Image-derived queries	Safety	Rouge/GPT-Eval	ASR/RR
EFUF (Xing et al., 2024)	Images from MSCOCO	Captions from MSCOCO	Hallucination	BLEU,Perplexity,CHAIR	BLEU,Perplexity,CHAIR

Table 2: Overview of existing benchmarks for LLM unlearning. Some (Łucki et al., 2024; Lynch et al., 2024) focus on adversarial evaluation using existing datasets. TR: Truth Ration, EM: Exact Match.

ine whether models can forget copyrighted material. TOFU (Maini et al., 2024) introduces synthetic and real author profiles to evaluate the removal of personal information and its impact on downstream utility. RWKU (Jin et al., 2024) focuses on public-figure knowledge, employing structured probing and adversarial evaluations—such as membership inference and jailbreak tests—to quantify unlearning effectiveness and robustness against knowledge recovery. MUSE (Shi et al., 2025) offers a comprehensive framework by integrating WHP and NEWS (Li et al., 2024d) to benchmark diverse unlearning strategies.

The WMDP dataset comprises expert-curated multiple-choice questions spanning biosecurity, cybersecurity, and chemical safety, serving as a benchmark for assessing **safety**-oriented unlearning (Li et al., 2024b). For **bias** mitigation, Dige et al. (2024) employ datasets such as BBQ (Parrish et al., 2022), StereoSet (Nadeem et al., 2021), Civil Comments (Duchêne et al., 2023), and Reddit-Bias (Barikeri et al., 2021), while Liu et al. (2025a) leverage CrowS-Pairs (Nangia et al., 2020) for similar evaluations. Furthermore, Liu et al. (2024d) investigate **hallucination** mitigation performance using TruthfulQA (Lin et al., 2022) and HaluEval (Li et al., 2023a).

Multimodal Benchmarks Similarly, datasets focusing on **privacy** have been developed most extensively. FIUBench (Ma et al., 2024b) pairs synthetic face images with generated private data to examine whether models can forget sensitive identity attributes while retaining general reasoning. MLLMU-Bench (Liu et al., 2024b) integrates fictitious and celebrity faces with related profiles, enabling multimodal evaluation of privacy unlearning across text and vision. Besides, MMUBench (Li et al., 2024a) explores common visual concepts—such as cartoon characters and *artistic styles*—paired with related knowledge to study the forgetting of visual–semantic associations. CLEAR (Dontsov et al., 2025) bridges textual and visual modalities by linking synthetic face images with text data from TOFU.

Beyond privacy, SAFEERASER (Chen et al., 2025) focuses on **safety** unlearning using a mixture of real and synthetic harmful images paired with image-derived queries to evaluate the removal of unsafe visual knowledge. It evaluates forgetting efficacy through attack success rate (ASR) and refusal rate (RR). For **hallucination**, EFUF (Xing et al., 2024) employs MSCOCO images and captions (Lin et al., 2014) to analyze how unlearning reduces visual hallucinations in captioning models. It adopts two specialized hallucination metrics—CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023b)—to quantify the extent of hallucination reduction and the improvement in factual grounding.

Takeaway 2: **i)** The majority emphasize privacy-related forgetting, leaving other domains under-represented. **ii)** Most benchmarks focus on textual and synthetic image data, while audio and video modalities remain unexplored.

4 LLM Unlearning Methods

4.1 Direct Fine-Tuning

Relabeling-based fine-tuning. This method replaces the original responses with neutral or generic alternatives and subsequently fine-tunes the model on the modified data to mitigate the effect of undesired content (Jin et al., 2024; Eldan et al., 2024).

Gradient ascent. Jang et al. (2023) apply gradient ascent (GA) on the next-token loss over the

forget set, equivalent to minimizing the negative log-likelihood:

$$\mathcal{L}_{GA}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[-\log\left(p(y \mid x;\theta)\right)\right]. \quad (2)$$

However, GA often degrades model quality, producing uniform, low-quality outputs. To mitigate this, Liu et al. (2022) add a gradient descent (GD) loss on the retain set $\mathcal{D}_r$ as regularization. KL Minimization (Yao et al., 2024) instead use KL divergence to align the fine-tuned model with the original on $\mathcal{D}_r$. Pan et al. (2025) propose MOLLM, a multi-objective optimization–based method that replaces the term $-\log(p)$ with its unlearning variant $-\log(1-p)$ to avoid gradient explosion.

Reinforcement learning. The Quark method (Lu et al., 2022) is a pioneering approach to applying reinforcement learning (RL) for LLM unlearning. It uses a reward model and Proximal Policy Optimization (PPO) (Schulman et al., 2017) to reduce undesirable behaviors such as toxicity, repetition, and unwanted sentiment. The reward model assesses the output quality using task-specific measures of toxicity and sentiment. Quark alternates between collecting samples, sorting them into reward-based quantiles labeled with reward tokens, and applying language modeling loss conditioned on these tokens, with a KL-divergence penalty to stay close to the original model. Kassem et al. (2023) proposed DeMem, which leverages a negative similarity reward. This approach trains the LLMs to develop a paraphrasing policy on the forget dataset, generating dissimilar tokens that minimize memorization while preserving semantic coherence.

Preference optimization. Preference optimization (PO) aligns model behavior with human-defined preferences by using pairwise comparisons or ranking data to guide outputs toward desired responses. Given a preference dataset $\mathcal{D}_p = \{(x_i, y_{i,w}, y_{i,l})\}_{i\in[n]}$, where $y_{i,w}$ and $y_{i,l}$ represent responses to input x_i, the preference $y_{i,w} > y_{i,l}$ is derived from human comparisons. Direct preference optimization (DPO) (Rafailov et al., 2023) minimizes the following objective function:

$$\mathcal{L}_{\text{DPO},\beta}(\theta) = -\frac{1}{\beta}\mathbb{E}_{\mathcal{D}_p}\left[\log\sigma\left(\beta\log\frac{p(y_w \mid x;\theta)}{p(y_w \mid x;\theta_{ori})} - \beta\log\frac{p(y_l \mid x;\theta)}{p(y_l \mid x;\theta_{ori})}\right)\right], \quad (3)$$

where σ is the sigmoid function and β is the inverse temperature controlling the preference strength. Maini et al. (2024) pioneered the application of DPO to unlearning by framing the forget set as a preference set. The original responses are denoted as y_l, while refusal responses, such as "*I do not know the answer*," are designated as y_w. Inspired by this, Zhang et al. (2024b) proposed negative preference optimization (NPO), a DPO variant that uses only negative responses from the forget set, disregarding y_w in Eq. (3):

$$\mathcal{L}_{\text{DPO},\beta}(\theta) = -\tfrac{2}{\beta}\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\log\sigma\left(-\beta\log\tfrac{p(y|x;\theta)}{p(y|x;\theta_{ori})}\right)\right]. \quad (4)$$

4.2 Localized Parameter Modification

Representation engineering. RMU (Li et al., 2024b) focuses on unlearning hazardous knowledge in LLMs by fine-tuning the lower layer l to redirect internal representations of token t in the forget set toward a fixed-noise vector $\mathbf{u}$:

$$\mathcal{L}_{\text{forget}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\frac{1}{L_x}\sum_{t\in x}\|\mathcal{H}_\theta^{(l)}(t) - c\cdot\mathbf{u}\|_2^2, \quad (5)$$

where L_x is the number of tokens in x and c is some hyperparameter that controls activation scaling, $\mathcal{H}_\theta^{(l)}(t)$ denotes the internal activations of token t at layer l. Simultaneously, it ensures that preserved knowledge remains consistent with the original model by aligning its representations, denoted as:

$$\mathcal{L}_{\text{retain}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_r}\frac{1}{L_x}\sum_{t\in x}\|\mathcal{H}_\theta^{(l)}(t) - \mathcal{H}_{\theta_{\text{ori}}}^{(l)}(t)\|_2^2. \quad (6)$$

Dang et al. (2025) proposed adaptive RMU, which dynamically scales the random unit vector $\mathbf{u}$ with an adaptive coefficient $\beta\|\mathcal{H}_{\theta_{\text{ori}}}^{(l)}(x)\|_2^2$ for improved unlearning across layers, unlike RMU's fixed-scaling coefficient c.

Locate-then-unlearn methods. This method focuses on identifying and localizing key model components (e.g., layers or neurons) that are critical for unlearning. DEPN (Wu et al., 2023) leverages privacy attribution via gradient integration to identify privacy-sensitive neurons. It quantifies each neuron's contribution to privacy leakage by efficiently approximating the integral of the gradient changes using a limited number of steps (e.g., $m = 20$). Specifically, the privacy attribution score $\text{Att}(w_k^l)$ for a neuron w_k^l at layer l is computed using the following cumulative gradient integration:

$$\text{Att}(w_k^l) = \int_0^{\beta_k^l}\frac{\partial p(y|x,\alpha_k^l)}{\partial w_k^l}d\alpha_k^l \approx \frac{\beta_k^l}{m}\sum_{j=1}^{m}\frac{\partial p(y|x,\frac{j}{m}\beta_k^l)}{\partial w_k^l}, \quad (7)$$

where β_k^l is the activation value of the neuron, α_k^l represents the modified activation value, and $p(y|x)$ is the conditional probability of the model predicting the private information.

WAGLE formulates unlearning as a bi-level optimization and uses weight attribution to measure how individual weights influence forget and retain losses. The resulting sensitivity score captures the trade-off between unlearning and utility, computed efficiently with a diagonal Hessian approximation (Jia et al., 2024). Guo et al. (2024) propose a mechanistic framework that localizes knowledge through causal probing or path patching, followed by targeted fine-tuning. Needle (Hong et al., 2024) further identifies and disrupts concept vectors in MLP layers that encode specific knowledge.

4.3 Leveraging Auxiliary Models

Contrastive decoding. ULD (Ji et al., 2024) trains an auxiliary model on the forget set to guide unlearning during decoding. The model aims to capture knowledge unique to the forget set while producing uniform predictions on the retain set. Its objective is formulated as the inverse of Eq. (1):

$$\min_{\theta_a} \mathcal{L}(\theta_a) = \min_{\theta_a} \{\mathcal{L}_f(\theta_a) - \beta \mathcal{L}_r(\theta_a)\}. \tag{8}$$

where $\mathcal{L}_r(\theta_a)$ is defined as the cross-entropy with a uniform distribution. For efficiency, ULD reuses the first k transformer layers and the language model head from the original LLM to generate logits over the vocabulary.

Knowledge distillation. RKLD (Wang et al., 2024) identifies tokens requiring unlearning by detecting those with consistently elevated logits after fine-tuning. The teacher model is defined as:

$$l(y|x;\theta_{ori}) - \alpha \cdot ReLU(l(y|x;\theta_a) - l(y|x;\theta_{ori})) \tag{9}$$

where $l(y|x;\theta_{ori})$ and $l(y|x;\theta_a)$ represent the logits of the original and the auxiliary model, respectively, and α is a hyperparameter that controls the forgetting strength, respectively. Moreover, Dong et al. (2024) introduced a self-distillation method that assumes tokens like named entities or nouns contain sensitive information requiring unlearning. Key tokens are identified via syntactic analysis, and the teacher's logits are obtained by subtracting one-hot vectors (scaled by a hyperparameter) for these tokens from the original logits, thereby enforcing selective forgetting.

Task vectors. Task vectors steer the model behavior by editing the weight space, supporting operations such as addition or negation for unlearning and multi-task learning. Negating a task vector can suppress specific behaviors, such as reducing toxic language generation (Ilharco et al., 2023). Liu et al. (2024d) further improved fine-tuning by introducing modules for guided distortion, random disassociation, and preservation divergence to target harmful knowledge. Ethos (Gao et al., 2024) separates general and undesired knowledge by projecting task vectors onto principal components. Dou et al. (2025) leverage weight saliency mapping to selectively update model parameters and supports task vector negation for effective and stable unlearning.

4.4 Input/Output-based Methods

Input/output-based unlearning methods offer flexibility by using prompt engineering and post-processing without modifying the model weights or architecture. Liu et al. (2024a) propose training a prompt classifier to identify prompts within the scope of unlearning and corrupt them in the embedding space using zeroth-order optimization. Thaker et al. (2024) show that simple guardrails like prompting and input/output filtering can effectively support unlearning independently or alongside fine-tuning. Pawelczyk et al. (2025) propose in-context unlearning by constructing tailored prompts, where the labels of the data to be forgotten are flipped. Bhaila et al. (2025) propose SPUL, which uses learnable soft prompts to make large language models forget specific training data without updating model parameters.

4.5 Multimodal Methods

Numerous studies have extended unlearning techniques originally developed for textual LLMs (e.g., GA, GD, NPO) to vision–language (Ma et al., 2024b; Liu et al., 2024b; Dontsov et al., 2025) and audio–language models (Liu, 2025). They convert textual data into multimodal inputs, typically keeping the non-text modality parameters frozen. In contrast, CLIPErase (Yang et al., 2025) introduces a forgetting module that minimizes the similarity between image and text embeddings in the forget set, actively disrupting cross-modal alignment to erase specific visual–textual associations. Li et al. (2025b) employ a sparse autoencoder trained on the vision encoder to suppress neurons associated with target visual concepts, enabling parameter-free and selective forgetting at inference. Liu et al.

(2025c) further propose modality-aware neuron pruning, which removes neurons contributing to target knowledge within or across modalities, addressing the limitation of prior approaches that forget under multimodal inputs but still "remember" when presented with unimodal ones.

Table 3 compares existing unlearning methods across three dimensions: *utility preservation*, *forgetting efficacy*, and *efficiency*. These methods embody different design principles and trade-offs. For example, GA excels in robustly forgetting target data but significantly compromises utility, often resulting in degraded model performance. Methods like PO and RL strike a more favorable balance, achieving high utility with moderate forgetting and efficiency, making them appealing for scenarios requiring minimal side effects. Locate-then-unlearn techniques stand out by offering both high utility preservation and strong forgetting capabilities, although they often involve computationally intensive attribution analysis. Approaches that rely on auxiliary models, such as contrastive decoding or knowledge distillation, tend to suffer from reduced efficiency due to added model complexity. Ultimately, the optimal choice depends on whether the goal emphasizes forgetting efficacy, utility retention, or computational efficiency.

Takeaway 3: There is a *consistent* trade-off among utility preservation, forgetting efficacy, and efficiency, and the optimal method depends on the specific application requirements.

Method	Utility Preservation	Forgetting Efficacy	Efficiency
Relabeling-Based Fine-Tuning	Low	Medium	High
Gradient Ascent	Low	High	High
Preference Optimization	High	Medium	High
Reinforcement Learning	High	Medium	High
Representation Engineering	High	Medium	Medium
Locate-Then-Unlearn Methods	High	High	Medium
Contrastive Decoding	Medium	Medium	Low
Task Vector	Low	Medium	Low
Knowledge Distillation	Medium	Medium	Low
Input/Output-based Unlearning	High	Low	High
Multimodal Methods	High	High	Medium

Table 3: Comparison of various methods in terms of utility preservation, robust forgetting, efficiency.

5 Future Directions

Balancing utility preservation and forgetting efficacy remains challenging. As discussed above, achieving effective forgetting without compromising model utility is inherently difficult. Moreover, forgetting locality further complicates the problem—removing knowledge about one entity (e.g., "Donald Trump") may inadvertently erase related information (e.g., "Ivanka Trump"). Moreover, many unlearning methods lack robustness, as forgotten content can often be recovered through adversarial or probing attacks (Jin et al., 2024; Shi et al., 2025). Ensuring both reliable forgetting and stable utility thus remains an open challenge.

Theoretical frameworks are methodologically demanding. Existing LLM unlearning methods often lack formal guarantees of effectiveness. While locate-then-unlearn approaches (Wu et al., 2023; Jia et al., 2024) enhance interpretability, they do not establish a rigorous theoretical foundation. A crucial future direction is to develop a comprehensive framework that formally defines and ensures its effectiveness. This could involve leveraging principles from information theory (Jeon et al., 2024) and other theoretical approaches to provide a more principled understanding of LLM unlearning (Koloskova et al., 2025; Zhang et al., 2024a).

Real-world complexity is crucial for robust evaluation. Most existing methods focus on forgetting specific data points and rely on explicitly provided target sequences. In practice, unlearning requests are often broader and less well-defined. Future research should address more complex scenarios, such as removing implicit or derived knowledge rather than only explicit facts. Real-world demands may also involve deleting all information related to a particular organization or entire domains of knowledge. To meet these challenges, unlearning methods must become composable, interpretable, and easy to deploy, while remaining adaptable to frequent model updates.

6 Conclusion

We provided a comprehensive survey of recent advances in LLM unlearning. We began by defining the problem and outlining the foundational settings of LLM unlearning. To offer a structured understanding, we proposed a novel taxonomy that categorizes existing research from diverse perspectives. We further explored the methodologies used to implement unlearning and evaluates the effectiveness of these approaches in achieving the desired forgetting. Finally, we examined the key challenges in the field and identified promising directions for future research, thus offering valuable insights for researchers and practitioners.

Limitations

This survey mainly has the following limitations:

No experimental benchmarks Without original experiments, this paper cannot offer empirical validation of the theories or concepts. This limits the paper's ability to contribute new, verified knowledge to the field.

Potential omissions We have made our best effort to compile the latest advancements. Given the rapid development of this field, it is possible that some recent work may have been missed.

Ethics and Broader Impact

We anticipate no significant ethical concerns in our work. As a survey of recent progress in this research area, our study does not involve experimental implementation, the use of sensitive datasets, or the employment of annotators for manual labeling.

References

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Karuna Bhaila, Minh-Hao Van, and Xintao Wu. 2025. Soft prompting for unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4046–4056, Albuquerque, New Mexico. Association for Computational Linguistics.

Hwan Chang and Hwanhee Lee. 2025. Which retain set matters for LLM unlearning? a case study on entity unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5966–5982, Vienna, Austria. Association for Computational Linguistics.

Junkai Chen, Zhijie Deng, Kening Zheng, Yibo Yan, Shuliang Liu, PeiJun Wu, Peijie Jiang, Jia Liu, and Xuming Hu. 2025. Safeeraser: Enhancing safety in multimodal large language models through multimodal machine unlearning. *arXiv preprint arXiv:2502.12520*.

Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. 2025. Opt-out: Investigating entity-level unlearning for large language models via optimal transport. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28280–28297, Vienna, Austria. Association for Computational Linguistics.

Council of the European Union. 2016. General data protection regulation (GDPR). Document 32016R0679.

Council of the European Union. 2024. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *Proposal for a regulation of the European parliament and of the council*.

Huu-Tien Dang, Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. 2025. On effects of steering latent representation for large language model unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23733–23742.

Omkar Dige, Diljot Arneja, Tsz Fung Yau, Qixuan Zhang, Mohammad Bolandraftar, Xiaodan Zhu, and Faiza Khattak. 2024. Can machine unlearning reduce social bias in language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 954–969.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. 2024. Unmemorization in large language models via self-distillation and deliberate imagination. *arXiv preprint arXiv:2402.10052*.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. 2025. CLEAR: Character unlearning in textual and visual modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20582–20603, Vienna, Austria. Association for Computational Linguistics.

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. 2025. Avoiding copyright infringement via large language model unlearning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5176–5200, Albuquerque, New Mexico. Association for Computational Linguistics.

Corentin Duchêne, Henri Jamet, Pierre Guillaume, and Réda Dehak. 2023. A benchmark for toxic comment classification on civil comments dataset. *CoRR*, abs/2301.11125.

Ronen Eldan, Mark Russinovich, and Mark Russinovich. 2024. Who's harry potter? approximate unlearning for LLMs.

Lei Gao, Yue Niu, Tingting Tang, Salman Avestimehr, and Murali Annavaram. 2024. Ethos: Rectifying language models in orthogonal parameter space. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2054–2068, Mexico City, Mexico. Association for Computational Linguistics.

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. 2024. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. *arXiv preprint arXiv:2410.12949*.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. 2024. Intrinsic evaluation of unlearning using parametric knowledge traces. *arXiv preprint arXiv:2406.11614*.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. 2024. Jogging the memory of unlearned llms through targeted relearning attacks. In *Neurips Safe Generative AI Workshop 2024*.

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

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

Dongjae Jeon, Wonje Jeung, Taeheon Kim, Albert No, and Jonghyun Choi. 2024. An information theoretic metric for evaluating unlearning models. *arXiv preprint arXiv:2405.17878*.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Rao Kompella, Sijia Liu, and Shiyu Chang. 2024. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. *arXiv preprint arXiv:2406.08607*.

Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. 2024. WAGLE: Strategic weight attribution for effective and modular unlearning in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Rwku: Benchmarking real-world knowledge unlearning for large language models. *arXiv preprint arXiv:2406.10890*.

Aly Kassem, Omar Mahmoud, and Sherif Saad. 2023. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4379, Singapore. Association for Computational Linguistics.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. 2025. Certified unlearning for neural networks. In *Forty-second International Conference on Machine Learning*.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. 2024a. Single image unlearning: Efficient machine unlearning in multimodal large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Rassin Lababidi, Alexandr Wang, and Dan Hendrycks. 2024b. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning*.

Qing Li, Jiahui Geng, Zongxiong Chen, Kun Song, Lei Ma, and Fakhri Karray. 2025a. Internal activation revision: Safeguarding vision language models without parameter update. *arXiv preprint arXiv:2501.16378*.

Qing Li, Jiahui Geng, Derui Zhu, Fengyu Cai, Chenyang Lyu, and Fakhri Karray. 2025b. Sauce: Selective concept unlearning in vision-language models with sparse autoencoders. *arXiv preprint arXiv:2503.14530*.

Qing Li, Chenyang Lyu, Jiahui Geng, Derui Zhu, Maxim Panov, and Fakhri Karray. 2024c. Reference-free hallucination detection for large vision-language models. *arXiv preprint arXiv:2408.05767*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024d. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18600–18607.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human

falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pages 243–254. PMLR.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024a. Large language model unlearning via embedding-corrupted prompts. *arXiv preprint arXiv:2406.07933*.

Dianqing Liu, Yi Liu, Guoqing Jin, and Zhendong Mao. 2025a. Mitigating biases in language models via bias unlearning. *arXiv preprint arXiv:2509.25673*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, and 1 others. 2025b. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14.

Zhe Liu. 2025. Unlearning llm-based speech recognition models. In *Proc. Interspeech 2025*, pages 3214–3218.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. 2024b. Protecting privacy in multimodal large language models with mllmu-bench. *arXiv preprint arXiv:2410.22108*.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024c. Machine unlearning in generative ai: A survey. *arXiv preprint arXiv:2407.20516*.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024d. Towards safer large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1817–1829, Bangkok, Thailand. Association for Computational Linguistics.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. 2025c. Modality-aware neuron pruning for unlearning in multimodal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5913–5933, Vienna, Austria. Association for Computational Linguistics.

Huimin Lu, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2024. Towards transfer unlearning: Empirical evidence of cross-domain bias mitigation. *arXiv preprint arXiv:2407.16951*.

Ximing Lu, Sean Welleck, Jack Hessel, and Yejin Choi. 2022. Quark: Controllable text generation with reinforced unlearning. In *Advances in Neural Information Processing Systems*, volume 35, pages 27591–27609. Curran Associates, Inc.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. 2024. An adversarial perspective on machine unlearning for ai safety. *arXiv preprint arXiv:2409.18025*.

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

Weitao Ma, Xiaocheng Feng, Weihong Zhong, Lei Huang, Yangfan Ye, Xiachong Feng, and Bing Qin. 2024a. Unveiling entity-level unlearning for large language models: A comprehensive analysis. *arXiv preprint arXiv:2406.15796*.

Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, Yejin Choi, and 1 others. 2024b. Benchmarking vision language model unlearning via fictitious facial identity dataset. *arXiv preprint arXiv:2411.03554*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. 2024. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Zibin Pan, Shuwen Zhang, Yuesheng Zheng, Chi Li, Yuheng Cheng, and Junhua Zhao. 2025. Multi-objective large language model unlearning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2025. In-context unlearning: language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Nicholas Pochinkov, Nandi Schoots, and Nandi Schoots. 2024. Dissecting language models: Machine unlearning via selective pruning. *arXiv preprint arXiv:2403.01267*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Atakan Seyitoğlu, Aleksei Kuvshinov, Leo Schwinn, and Stephan Günnemann. 2024. Extracting unlearned information from llms with activation steering. *arXiv preprint arXiv:2411.02631*.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. 2025. MUSE: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*.

Pratiksha Thaker, Yash Maurya, Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. 2024. Guardrail baselines for unlearning in llms. *arXiv preprint arXiv:2403.03329*.

Bichen Wang, Yuzhe Zi, Yixin Sun, Yanyan Zhao, and Bing Qin. 2024. Rkld: Reverse kl-divergence-based knowledge distillation for unlearning personal information in large language models. *arXiv preprint arXiv:2406.01983*.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. DEPN: Detecting and editing privacy neurons in pretrained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886, Singapore. Association for Computational Linguistics.

Abudukelimu Wuerkaixi, Qizhou Wang, Sen Cui, Wutong Xu, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. 2025. Adaptive localization of knowledge negation for continual llm unlearning. In *Forty-second International Conference on Machine Learning*.

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2024. EFUF: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1181, Miami, Florida, USA. Association for Computational Linguistics.

Yi Xu. 2024. Machine unlearning for traditional models and large language models: A short survey. *arXiv preprint arXiv:2404.01206*.

Tomoya Yamashita, Yuuki Yamanaka, Masanori Yamada, Takayuki Miura, Toshiki Shibahara, and Tomoharu Iwata. 2025. Concept unlearning in large language models via self-constructed knowledge triplets. *arXiv preprint arXiv:2509.15621*.

Tianyu Yang, Lisen Dai, Xiangqi Wang, Minhao Cheng, Yapeng Tian, and Xiangliang Zhang. 2025. CLIPErase: Efficient unlearning of visual-textual associations in CLIP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30438–30452, Vienna, Austria. Association for Computational Linguistics.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand. Association for Computational Linguistics.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. 2024. A closer look at machine unlearning for large language models. *arXiv preprint arXiv:2410.08109*.

Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. 2024a. Towards certified unlearning for deep neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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

Xiao Zhang and Ji Wu. 2024. Dissecting learning and forgetting in language model finetuning. In *The Twelfth International Conference on Learning Representations*.

Zhexin Zhang, Junxiao Yang, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang.

2024c. Safe unlearning: A surprisingly effective and generalizable solution to defend against jailbreak attacks. *arXiv preprint arXiv:2407.02855*.

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