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

Anonymous ACL submission

## Abstract

This study investigates the machine unlearning techniques within the context of large language models (LLMs), referred to as *LLM unlearning*. LLM unlearning offers a principled approach to removing the influence of undesirable data (e.g., sensitive or illegal information) from LLMs, while preserving their overall utility without requiring full retraining. Despite growing research interest, there is no comprehensive survey that systematically organizes existing work and distills key insights; here, we aim to bridge this gap. We begin by introducing the definition and the paradigms of LLM unlearning, followed by a comprehensive taxonomy of existing unlearning studies. Next, we categorize current unlearning approaches, summarizing their strengths and limitations. Additionally, we review evaluation measures and benchmarks, providing a structured overview of current assessment methodologies. Finally, we outline promising directions for future research, highlighting key challenges and opportunities in the field.

## 1 Introduction

The widespread adoption of large language models (LLMs) 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., 2024b; Yao et al., 2024). 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 but 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" and require mechanisms to delete specific data upon request.

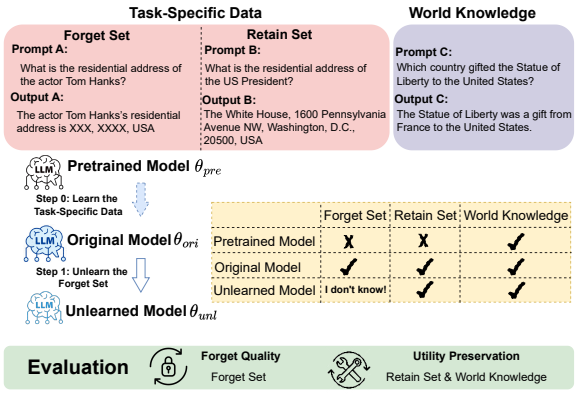


Figure 1: **Overview of LLM Unlearning.** LLM unlearning focuses on removing specific data (forget set) while minimizing the impact on related knowledge (retain set) and general world knowledge.

To address these challenges, the field of LLM unlearning has emerged, focusing on removing specific information or behaviors from models while preserving their overall performance. However, LLM unlearning faces significant technical challenges. One of the most pressing issues is the prohibitively high cost of retraining. Traditionally, 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 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 complexity of LLMs, with their millions, or even hundreds of billions of parame-

| Related Work | Date | Taxonomy | # Methodologies | Modality | Adv Evaluation |
|---|---|---|---|---|---|
| Si et al. (2023) | December 2023 | Yes | 3 | Unimodal | 0 |
| Xu (2024) | April 2024 | Yes | 3 | Unimodal | 0 |
| Liu et al. (2025) | June 2024 | No | 6 | Unimodal,Multimodal | 5 |
| Liu et al. (2024c) | July 2024 | No | 5 | Unimodal,Multimodal | 0 |
| Ours | May 2025 | Yes | 10 | Unimodal,Multimodal | 7 |

Table 1: Comparison of different surveys on the LLM Unlearning. *Adv Evaluation* refers to Adversarial Evaluation.

ters, further complicates the task of removing specific information without causing unintended side effects, such as performance degradation or catastrophic forgetting (Zhang et al., 2024a).

Table 1 provides a comparative overview of existing survey papers on LLM unlearning alongside our work. Most prior surveys were published before July 2024 and therefore do not capture recent advancements in the field. Moreover, these surveys either lack a systematic taxonomy or are limited to unimodal unlearning methods. Here we aim to bridge this gap. In particular, we offer a thorough overview of the field, including various unlearning and evaluation methods, and we make the following contribution

- We formalize the LLM unlearning paradigms and propose a comprehensive taxonomy to categorize the existing approaches. This taxonomy not only offers a structured understanding of the research landscape, but also helps researchers identify their areas of interest.
- We systematically review the existing methods, analyzing their strengths and weaknesses. We further examine the existing evaluation measures and benchmarks, highlighting the challenges of balancing utility preservation with forgetting quality.
- We discuss future research opportunities for LLM unlearning, including extending techniques to multimodal models and addressing complex real-world unlearning requests. These avenues aim to advance the field and address emerging challenges.

## 2 Preliminaries and Taxonomy

### 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 $\theta$ 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.

### 2.2 LLM Unlearning Paradigms

LLM unlearning follows two main paradigms. The *fine-tuning-then-unlearning* paradigm focuses on eliminating knowledge introduced during *fine-tuning*. As illustrated in Figure 1, this paradigm typically leverages synthetic data (e.g., TOFU (Maini et al., 2024) and FIUBench (Ma et al., 2024)) and partitions a task-specific dataset into a forget set $\mathcal{D}_f$ and a retain set $\mathcal{D}_r$ to highlight the unlearning precision. The original model $\theta_{ori}$ is first obtained by fine-tuning a pretrained model $\theta_{pre}$ on the task-specific data to encode the target knowledge. Unlearning techniques are then applied to reduce the model's reliance on the forget set, resulting in the unlearned model $\theta_{unl}$. The *direct-unlearning* paradigm focuses on eliminating knowledge acquired during the *pretraining* stage of $\theta_{ori}$, assuming the target knowledge originates from multiple points within the pretraining dataset. The forget and retain sets are typically sampled from pretraining corpora (Yao et al., 2024) or publicly available datasets such as Wiki (Jin et al., 2024). This paradigm is also extensively applied in safety alignment tasks, where it aims to eradicate hazardous knowledge and to mitigate risks such as misuse or jailbreak attacks (Li et al., 2024b; Zhang et al., 2024b).

### 2.3 Taxonomy

We present a comprehensive taxonomy of LLM unlearning in Figure 2, outlining existing research

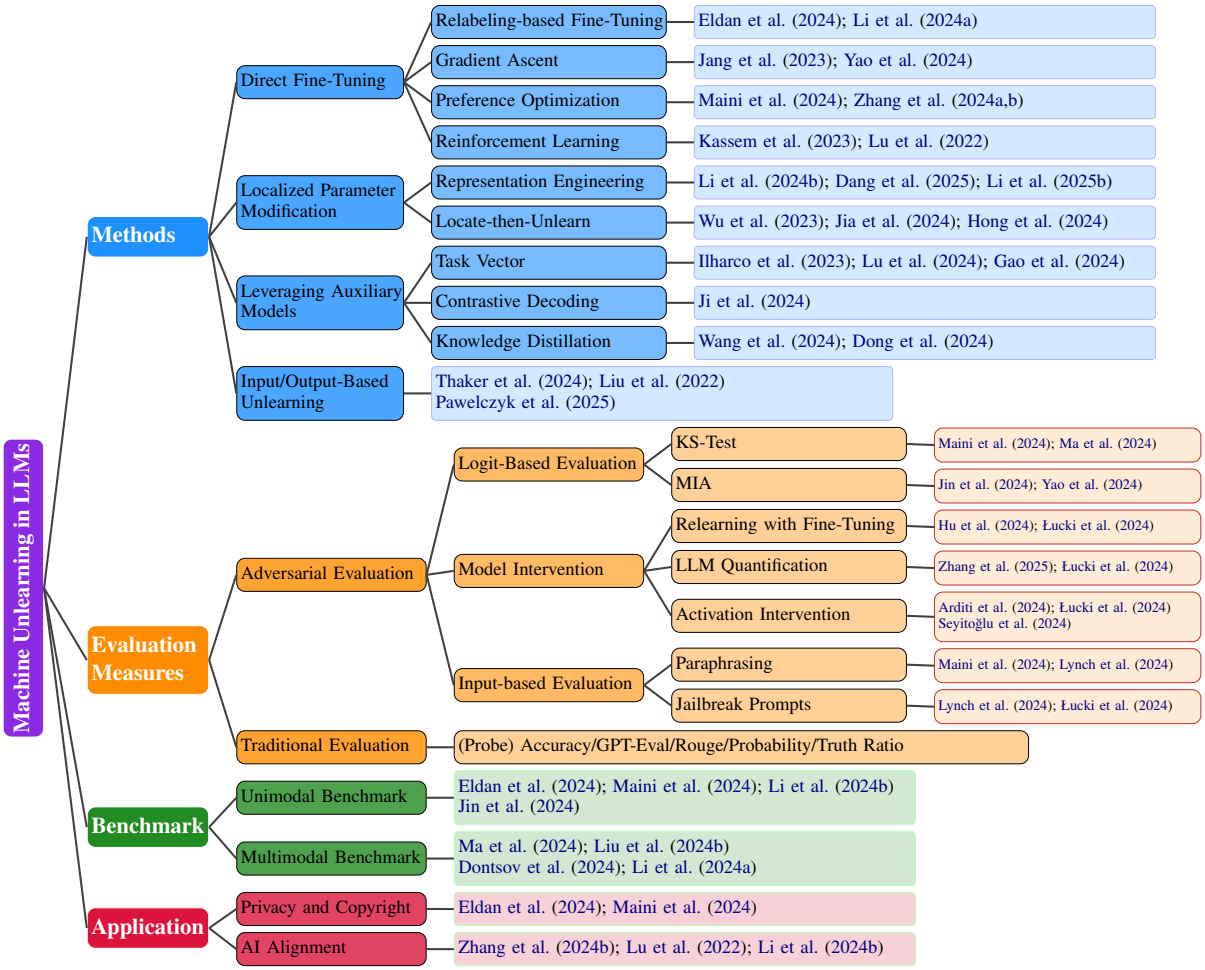


Figure 2: The taxonomy of machine unlearning in LLMs.

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 LLM Unlearning Methods

### 3.1 Direct Fine-Tuning

**Relabeling-based fine-tuning** first replaces the original responses with generic or neutral substitutes, as present above. The LLM is then fine-tuned on relabeled data to reduce the effect of undesired information (Jin et al., 2024; Eldan et al., 2024).

**Gradient ascent (GA)** 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. Yao et al. [2024] instead use KL divergence to align the fine-tuned model with the original on $\mathcal{D}_r$:

$$\mathcal{L}_{KL}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_r}\left[D_{KL}\left(p(y \mid x;\theta_{ori}) \,\|\, p(y \mid x;\theta)\right)\right]. \quad (3)$$

These regularization techniques can also be applied in other unlearning methods to preserve the utility.

**Reinforcement learning (RL)** The Quark method (Lu et al., 2022) is a pioneering approach

to applying reinforcement learning 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 (PO)** was first designed to align model behavior to human-defined preferences. Specifically, it leverages pairwise comparisons or ranking data to guide the model toward producing outputs that best match desired preferences. 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:

$$\mathcal{L}_{\text{DPO},\beta}(\theta) = -\frac{1}{\beta}\mathbb{E}_{\mathcal{D}_p}\left[\log\sigma\left(\beta\log\frac{p(y_w \mid x;\theta)}{p(y_w \mid x;\theta_{ori})} - \beta\log\frac{p(y_l \mid x;\theta)}{p(y_l \mid x;\theta_{ori})}\right)\right], \tag{4}$$

where $\sigma$ is the sigmoid function and $\beta$ 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 $y_w$. This formulation guides the unlearning process by aligning the model's behavior with the preferred alternative responses. Inspired by this idea, Zhang et al. [2024a] proposed negative preference optimization (NPO), a DPO variant that uses only negative responses from the forget set, disregarding $y_w$ in Eq. (4):

$$\mathcal{L}_{\text{DPO},\beta}(\theta) = -\frac{2}{\beta}\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\log\sigma\left(-\beta\log\frac{p(y|x;\theta)}{p(y|x;\theta_{ori})}\right)\right]. \tag{5}$$

Zhang et al. [2024a] further theoretically showed that NPO converges to GA as $\beta \to 0$ and the speed toward collapse using NPO is exponentially slower than GA.

### 3.2 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 $\mathbf{u}$:

$$\mathcal{L}_{\text{forget}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\frac{1}{L_x}\sum_{t\in x}\|\mathcal{H}_\theta^{(l)}(t) - c\cdot\mathbf{u}\|_2^2, \tag{6}$$

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:

$$\mathcal{L}_{\text{retain}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_r}\frac{1}{L_x}\sum_{t\in x}\|\mathcal{H}_\theta^{(l)}(t) - \mathcal{H}_{\theta_{\text{ori}}}^{(l)}(t)\|_2^2. \tag{7}$$

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:

$$\text{Att}(w_k^l) = \int_0^{\beta_k^l}\frac{\partial p(y|x,\alpha_k^l)}{\partial w_k^l}d\alpha_k^l \approx \frac{\beta_k^l}{m}\sum_{j=1}^{m}\frac{\partial p(y|x,\frac{j}{m}\beta_k^l)}{\partial w_k^l}, \tag{8}$$

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

WAGLE (Jia et al., 2024) uses bi-level optimization to examine the interaction between weight adjustment and unlearning efficacy. By leveraging weight attribution, it quantifies the relationship between weight influence and the impact of forgotten

or retained data on LLM outputs. The unlearning sensitivity score for weight perturbation is obtained from the forget loss $S_i = \mathcal{L}_f(\epsilon \odot \theta(\epsilon)) - \mathcal{L}_f(\theta(1))$, where $\epsilon \odot \theta(\epsilon)$ is the weight-adjusted model, $\epsilon$ represents weight modifications, and $\epsilon = 1$ indicates no interference. The weights $\theta(\epsilon)$ minimize the retain loss $\mathcal{L}_r$ to preserve utility. WAGLE uses a diagonal Hessian approximation for computational efficiency and accuracy, with the sensitivity score expressed as:

$$S_i \propto [\theta]_i [\nabla \mathcal{L}_f(\theta)]_i - \frac{1}{\gamma} [\nabla \mathcal{L}_r(\theta)]_i [\nabla \mathcal{L}_f(\theta)]_i, \quad (9)$$

where $[\theta]_i$ is the $i$-th weight, $[\nabla \mathcal{L}_f(\theta)]_i$ and $[\nabla \mathcal{L}_r(\theta)]_i$ are the gradients of the forget and the retain losses for the $i$-th weight, respectively, and $\gamma$ is the Hessian parameter.

Beyond DEPN and WAGLE, Guo et al. [2024] introduced a mechanistic unlearning framework that combines fact lookup localization, using logistic regression probes or path patching to assess causal importance, with localized fine-tuning. Similarly, Needle (Hong et al., 2024) identifies and disrupts concept vectors in MLP layers encoding specific knowledge using vocabulary projections and causal tests.

### 3.3 Leveraging Auxiliary Models

These methods typically fine-tune an assistant model $\theta_a$ to replicate knowledge from $\mathcal{D}_f$. Its outputs are then used to adjust the original model's responses, mitigating the influence of $\mathcal{D}_f$ through the auxiliary model's weights or logits.

**Contrastive decoding** ULD (Ji et al., 2024) uses an auxiliary model trained on the forget set to guide the unlearning process during decoding. It claims that the auxiliary LLM should exclusively capture the unique knowledge within the forget set while preventing the retention of any information meant to be preserved. Ideally, this results in a uniform distribution over the retain set. The optimization objective of the auxiliary model is formulated as the inverse of Eq. (1):

$$\min_{\theta_a} \mathcal{L}(\theta_a) = \min_{\theta_a} \{\mathcal{L}_f(\theta_a) - \beta \mathcal{L}_r(\theta_a)\}. \quad (10)$$

The retain loss $\mathcal{L}_r(\theta_a)$ is specifically formulated as the cross-entropy with respect to the uniform distribution. To enhance the efficiency in the auxiliary model's implementation, the first $k$ transformer layers of the original LLM are reused, along with the language model head, to map hidden representations to output logits across the entire vocabulary.

**Knowledge distillation** uses a specialized unlearning teacher model to guide the unlearning process, providing signals to the student model to adjust the logits and to selectively forget specific information. RKLD (Wang et al., 2024) first identifies tokens requiring unlearning by detecting such with consistently increased logit values after fine-tuning. The formula for the unlearning teacher model is as follows:

$$l(y|x;\theta_{ori}) - \alpha \cdot ReLU(l(y|x;\theta_a) - l(y|x;\theta_{ori})) \quad (11)$$

where $l(y|x;\theta_{ori})$ and $l(y|x;\theta_a)$ represent the logits of the original and the auxiliary model, respectively, and $\alpha$ is a hyperparameter that controls the forgetting strength, respectively. This strategy offers more precise guidance for intentional forgetting while safeguarding other information. Moreover, Dong et al. [2024] introduced a self-distillation method that assumes tokens like named entities or nouns contain sensitive information requiring unlearning. To identify these key tokens, they used a syntactic analysis tool for extraction. The teacher model's logits were derived by subtracting hyperparameter-controlled one-hot vectors for the key tokens from the logits of the original model.

**Task vectors** , defined as $\tau = \theta_a - \theta_{ori}$, steer the model behavior by editing the weight space, thus enabling operations such as negation and addition for applications such as unlearning and multi-task learning. Negating task vectors effectively suppress behaviors such as mitigating toxic language generation (Ilharco et al., 2023). Liu et al. [2024d] enhanced fine-tuning with modules targeting harmful knowledge: guided distortion, random disassociation, and preservation divergence. Ethos (Gao et al., 2024) distinguishes general and undesired knowledge by projecting task vectors onto principal components. By negating only undesired components, Ethos minimizes collateral damage to model utility, thus achieving unlearning.

### 3.4 Input/Output-based Unlearning

Input/output-based unlearning methods offer flexibility by using prompt engineering and post-processing without modifying the model weights or the architecture. Liu et al. [2024a] proposed training a prompt classifier to identify prompts within the scope of unlearning and efficiently corrupting them in the embedding space using zeroth-order optimization. Thaker et al. [2024] showed that simple

| Method | Utility Preservation | Robust Forgetting | 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 |

Table 2: Comparison of various methods in terms of utility preservation, robust forgetting, efficiency.

guardrails like prompting and input/output filtering can effectively support unlearning independently or alongside fine-tuning. Pawelczyk et al. [2025] proposed in-context unlearning by constructing tailored prompts, where the labels of the data to be forgotten are flipped.

**Summary of Unlearning Methodologies** Table 2 provides a comparative overview of existing unlearning methodologies, assessing their strengths and limitations across three key dimensions: utility preservation, robust forgetting, and efficiency. These methods reflect varying design principles, leading to distinct trade-offs. For example, GA excels in robustly forgetting target data but significantly compromises utility, often resulting in degraded model performance. On the other hand, 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. Overall, the choice of method hinges on the specific unlearning goal—whether prioritizing forgetting effectiveness, maintaining model utility, or optimizing computational cost.

## 4 Benchmarks

This section provides a detailed description of the commonly used benchmarks, addressing areas such as copyright, privacy, and AI alignment. Notably, WMDP and RWKU are designed for direct-unlearning, while other benchmarks are used within the fine-tuning-then-unlearning paradigm.

### 4.1 Unimodal Benchmarks

**Who is Harry Potter (WHP)** (Eldan et al., 2024) evaluates unlearning of Harry Potter-related information using a dataset combining the original books (2.1M tokens) with synthetic content (1M tokens). Unlearning effectiveness is assessed through 330 Harry Potter-related questions scored with a GPT-4-based evaluation.

**TOFU** (Maini et al., 2024) includes 200 synthetic author profiles with 20 question-answer examples each, ensuring no overlap with existing training data. The benchmark also includes 100 real-world author profiles and 117 world facts, comprehensively evaluating model utility after unlearning.

**WMDP** (Li et al., 2024b) comprises 3,668 multiple-choice questions on biosecurity, cybersecurity, and chemical security, curated by experts to evaluate hazardous knowledge while excluding sensitive information. It serves as a benchmark for assessing LLMs' hazardous knowledge and unlearning methods for AI alignment.

**RWKU** (Jin et al., 2024) includes 200 unlearning targets centered on well-known public figures, comprising 13,131 multi-level forget probes and 11,379 neighbor probes. In addition, it incorporates a wide range of adversarial evaluation methods, including membership inference attacks (MIA), jailbreak prompts, and others.

### 4.2 Multimodal Benchmarks

**FIUBench** (Ma et al., 2024) comprises 400 synthetic faces paired with fictitious private data such as personal backgrounds, health records, criminal histories, phone numbers, occupations and incomes are randomly assigned. GPT-4o generates 20 question-answer pairs for each profile.

**MLLMU-Bench** (Liu et al., 2024b) contains 500 fictitious profiles and 153 celebrity profiles, each with 14+ question-answer pairs evaluated in multimodal (image+text) and unimodal (text-only) settings. It features 20.7k questions, with fictitious profiles generated by GPT-4o and real celebrity profiles reviewed by experts. The test set includes 3.5k paraphrased questions and 500 modified images with pose variations. An additional utility set uses celebrity profiles.

**CLEAR** (Dontsov et al., 2024) focuses on person unlearning, characterizing unlearning across textual and visual modalities. It generates consistent images through a comprehensive strategy

| Benchmark | Data Source | | # Volume | Traditional Evaluation | Adversarial Evaluation |
|---|---|---|---|---|---|
| | Image | Text | | | |
| WHP (Eldan et al., 2024) | - | Harry Potter books | 330 | GPT-Eval/Probability | - |
| TOFU (Maini et al., 2024) | - | Fictitious Profiles | 4,000 | Probability/Rouge/TR | Paraphrasing |
| WMDP (Li et al., 2024b) | - | Safety-related documents | 3,668 | Probe Accuracy | Jailbreak Prompts |
| RWKU (Jin et al., 2024) | - | Public figure knowledge | 2,4510 | Rouge | Paraphrasing, Jailbreak Prompts,MIA |
| (Łucki et al., 2024) | - | WMDP | - | Probe Accuracy | Relearning,Intervention,Jailbreak Prompts,Pruning |
| (Lynch et al., 2024) | - | WHP | - | GPT-Eval | Relearning, Jailbreak Prompts |
| FIUBench (Ma et al., 2024) | Synthetic face images | Generated private data | 8,000 | Rouge/GPT-Eval/TR/EM | - |
| MMUBench (Li et al., 2024a) | Common visual concepts | Related knowledge | 2,000 | Rouge/GPT-Eval/EM | Paraphrasing, MIA, Jailbreak Prompts |
| MLLMU-Bench (Liu et al., 2024b) | Fictitious and celebrity faces | Related profiles | 20,754 | Accuracy/Rouge | Paraphrasing |
| CLEAR (Dontsov et al., 2024) | Synthetic face images | Data from TOFU | 7,700 | TR, Probability, Rouge | Paraphrasing |

Table 3: 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.

and links them to the corresponding author-related questions from TOFU. It includes a total of 200 fictitious individuals linked with 3.7k visual question-answer pairs and 4k textual question-answer pairs, enabling a thorough evaluation of unimodal and multi-modal unlearning techniques.

## 5 Evaluation Measures

We group existing measures into two categories. *Classical evaluation* uses standard utility metrics to assess forgetting (via performance drop on the forget set) and utility preservation (on the retain set and world knowledge). *Adversarial evaluation* tests whether forgetting is robust or merely superficial suppression.

### 5.1 Classical Evaluation

To evaluate utility preservation and forgetting effectiveness, several classical metrics are commonly employed. *(Probing) Accuracy* assesses whether the unlearned model maintains its world knowledge without performance degradation. *GPT-Eval* uses large language models as evaluators to measure multiple dimensions beyond traditional metrics; for instance, Ma et al. (2024) employ GPT-4o Mini to score correctness, helpfulness, and relevance, producing an overall score between 0 and 1. *ROUGE* quantifies the similarity between generated outputs and ground truth responses (Maini et al., 2024; Yuan et al., 2024). *Probability-based* evaluation estimates the model's confidence by computing the normalized conditional likelihood of the target output. *Truth Ratio* is specially proposed to compare the model's likelihoods of correct versus incorrect answers for a given question and has been used in many benchmarks (Maini et al., 2024; Ma et al., 2024; Dontsov et al., 2024). Specifically, it is the ratio of the average normalized conditional probability of perturbed incorrect answers $\bar{y} \in \mathcal{A}$ to that of a paraphrased correct answer $\tilde{y}$. Both $\bar{y}$ and $\tilde{y}$ can be generated by LLMs. A lower truth ratio indicates better forgetting. It is defined as:

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

On the retain and world knowledge sets, $\max(0, 1 - R_{\text{truth}}(y))$ is used to measure how well the model preserves relevant information.

### 5.2 Adversarial Evaluation

Recent work (Lynch et al., 2024; Ma et al., 2024) emphasize the essence of distinguishing real forgetting from mere suppression, where suppressed knowledge can be recovered via adversarial techniques. This motivates the summation of *adversarial evaluation* to capture such subtle distinctions.

**Input-based evaluation** assesses whether a model has truly forgotten information by modifying the input rather than probing its internal representations. *Paraphrasing* involves rephrasing questions or translating them into other languages (e.g., Spanish or Russian) 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 or 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 methods (e.g., optimized prefixes) to bypass unlearning and extract suppressed information (Łucki et al., 2024).

**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., 2024). Moreover, *Membership Inference Attacks (MIAs)* determine whether a specific data point is part of a model's training data (member) or originates from outside the training set (non-member). Hence, it is applied to evaluate unlearning efficacy (Jin et al., 2024; Yao et al., 2024). Jin et al. [2024] used various MIA methods to assess the robustness of unlearning and found that many unlearning methods failed under such attacks.

**Model intervention methods** evaluate the robustness of unlearning by directly modifying the model's parameters, activations, or numerical precision—interventions that may inadvertently reverse unlearning and expose residual memorization. *Relearning through fine-tuning* refers to the phenomenon where continued training on benign and loosely related data (e.g., "What is avian influenza?") causes the model to recover previously forgotten knowledge. This suggests that unlearning may suppress rather than fully remove the underlying representations (Hu et al., 2024; Łucki et al., 2024). *LLM quantization* offers another perspective, where reducing the model's precision—such as converting weights to 4-bit—can increase the likelihood of forgotten content re-emerging, thereby weakening the unlearning effect (Zhang et al., 2025; Łucki et al., 2024). Similarly, Łucki et al. (2024) evaluate neuron pruning as a means of assessing residual memorization. Finally, *activation intervention* techniques analyze the model's internal activations to identify and remove the so-called refusal direction—a vector derived from differences between the original and unlearned models. Suppressing this direction reduces refusal behavior and enables the model to regenerate responses that were assumed to be forgotten (Arditi et al., 2024; Łucki et al., 2024; Seyitoğlu et al., 2024; Li et al., 2025b).

## 6 Future Directions

**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.

**Multimodal unlearning shows promising potential.** While numerous multimodal datasets have been introduced for multimodal unlearning, current methods remain largely confined to text-based unlearning approaches (Ma et al., 2024; Liu et al., 2024b; Dontsov et al., 2024). Future research should prioritize the development of techniques capable of identifying and isolating modality-specific representations within MLLMs. Moreover, robust evaluation benchmarks are essential for assessing the effectiveness of multimodal unlearning methods in disentangling representations where knowledge is intertwined across both texts and images.

**Real-world complexity is crucial for robust evaluation.** Current unlearning methods primarily focus on removing specific data points from the model, requiring explicit target data points (sequences) to be provided. However, real-world unlearning requests may differ from this assumption. A significant future direction for LLM unlearning lies in addressing more complex requests, such as entity-level unlearning, which aims to remove all knowledge related to a specific entity across diverse contexts and associations. This involves not only forgetting explicit facts but also erasing implicit or derived knowledge. Choi et al. [2024] introduced datasets to evaluate the effectiveness of algorithms in entity-level unlearning tasks. Looking ahead, even more complex scenarios may emerge, such as removing all information about a specific organization, or erasing entire domains of knowledge, such as medical or criminal records.

## 7 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. Due to the rapid development in this field, there is still a possibility that some important work may have been overlooked.

## 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*.

Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. 2024. Opt-out: Investigating entity-level unlearning for large language models via optimal transport.

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.

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 Y Rogov, Ivan Oseledets, and Elena Tutubalina. 2024. Clear: Character unlearning in textual and visual modalities. *arXiv preprint arXiv:2410.18057*.

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.

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*.

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*.

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*.

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

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.

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*.

Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, Yejin Choi, and 1 others. 2024. 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*.

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.

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*.

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*.

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.

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

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*.

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

Zhexin Zhang, Junxiao Yang, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang. 2024b. 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*.