Human-Inspired Obfuscation for Model Unlearning: Local and Global Strategies with Hyperbolic Representations

Zekun Wang[1], **Jingjie Zeng**[1], **Yingxu Li**[1], **Liang Yang**[1,2]✉*, **Hongfei Lin**[1]
[1] School of Computer Science and Technology, Dalian University of Technology, China
[2] Key Laboratory of Social Computing and Cognitive Intelligence, Ministry of Education, China
zk_wang@mail.dlut.edu.cn,✉liang@dlut.edu.cn

Abstract

Large language models (LLMs) achieve remarkable performance across various domains, largely due to training on massive datasets. However, this also raises growing concerns over the exposure of sensitive and private information, making model unlearning increasingly critical. However, existing methods often struggle to balance effective forgetting with maintaining model utility. In this work, we propose **HyperUnlearn**, a human-inspired unlearning framework. We construct two types of fuzzy data **local** and **global** to simulate forgetting, and represent them in hyperbolic and Euclidean spaces, respectively. Unlearning is performed on a model with frozen early layers to isolate forgetting and preserve useful knowledge. Experiments demonstrate that **HyperUnlearn** effectively forgets sensitive content while maintaining the models language understanding, fluency, and benchmark performance, offering a practical trade-off between forgetting and capability preservation.

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; Qin et al., 2023; Chowdhery et al., 2023; Touvron et al., 2023b; Zhao et al., 2025; Azaria et al., 2024) have rapidly ascended to a position of prominence, demonstrating remarkable capabilities across a diverse array of natural language processing tasks and revolutionizing various application domains (Ouyang et al., 2022; Kojima et al., 2022; Radford et al., 2019; Lewkowycz et al., 2022; Roziere et al., 2023). One key factor behind their outstanding performance is that they are trained on massive datasets. (Hoffmann et al., 2022; Webson and Pavlick, 2021; Min et al., 2022; Liang et al., 2022). However, these massive datasets may contain privacy-sensitive information, leading to potential privacy breaches.

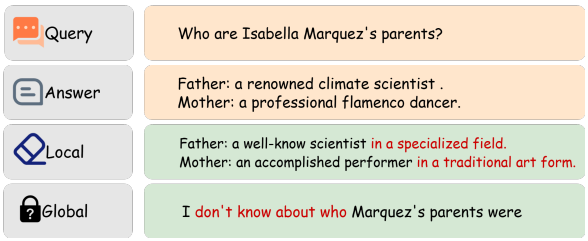

Figure 1: Taking inspiration from how humans forget, we use two types of obfuscation on the forget set to see which method better supports model unlearning.Here, **local forget** refers to data with localized obfuscation, while **global forget** denotes data that has been obfuscated in a more holistic or widespread manner.

Therefore, the task of model forgetting has become increasingly important. Consequently, the field of model unlearningwhich aims to selectively remove the influence of specific data points from a trained model without the prohibitive cost of retraining from scratchhas emerged as a critical area of research to address these pressing privacy challenges(Naing and Udomwong, 2024; Hua et al., 2024).

Current research in model unlearning predominantly focuses on post-training forgetting strategies (Liu et al., 2023; Chundawat et al., 2023; Jia et al., 2023; Zhang et al., 2024a; Zhao et al., 2024), many of which are centered around gradient-based manipulations (Jang et al., 2023). Despite their ingenuity, existing approaches often encounter two significant hurdles. Firstly, current research overlooks the impact of the intrinsic characteristics of the data during the forgetting process. Secondly, Forgetting can significantly degrade the original performance of the model (Yao et al., 2023), thereby limiting practical applicability.

Unlike traditional methods, we propose **HyperUnlearn**, a novel framework addressing key limitations in model unlearning through three complementary components. First, inspired by human forgetting (Wixted, 2004)which occurs either locally

* Corresponding Author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14354–14366
November 4-9, 2025

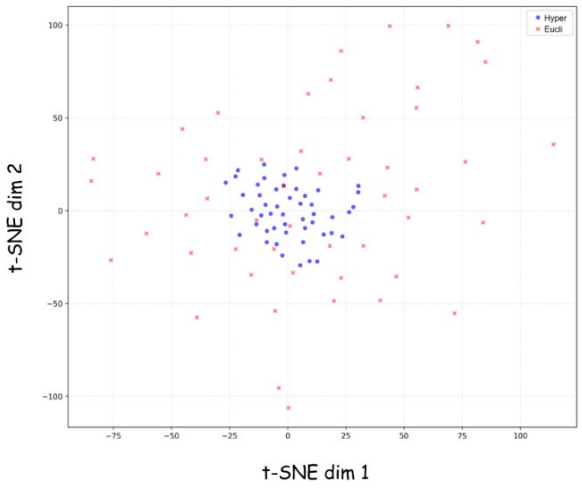

Figure 2: t-SNE Visualization of Embeddings in Euclidean and Hyperbolic Spaces.

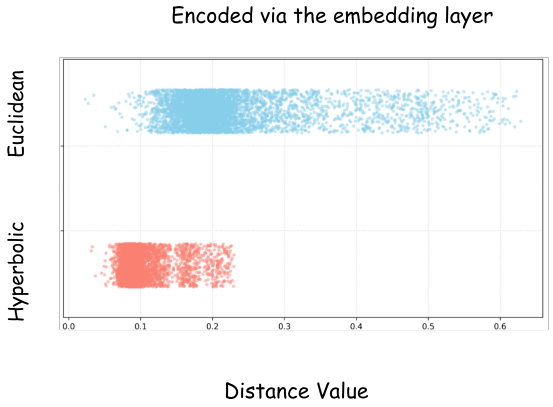

Figure 3: To examine the effect of representation space on the semantic distribution of the *Unlearn* dataset, we extract embeddings using the pre-trained `LLaMA2-7B` model and then project these embeddings into two different geometric spaces: (1) the original Euclidean space and (2) a hyperbolic space using Lorentz model.

(losing details while retaining the gist) or globally (complete loss of memory)we categorize the data to be forgotten into *local* and *global* fuzzy sets. These two types are handled differently during unlearning. Figure 1 illustrates examples of both forgetting strategies. Second, we address the semantic dispersion of unlearnable data. Performing gradient ascent in Euclidean space often yields unstable behavior. To counter this, we project unlearn data into hyperbolic space, which naturally induces a more compact distribution. As shown in Figure 2 and Figure 3, this tighter clustering facilitates more precise and controlled unlearning. Finally, to mitigate the issue of spurious unlearning (Zheng et al., 2025), we freeze the model's forward layers during training. This prevents the re-alignment of previously unlearned content and helps preserve the models core capabilities.

The integrated design of these modules within hyperUnlearn is specifically tailored to overcome the identified challenges in contemporary model unlearning. Each component serves a distinct purpose, collectively enabling a superior unlearning efficacy while concurrently safeguarding the model's original performance. The primary contributions of this paper are threefold:

1. Inspired by human cognition, we introduce *local* and *global forgetting* into model unlearning through two types of fuzzy data, enabling comparison of their effectiveness.

2. We present **HyperUnlearn**, a framework that integrates hyperbolic and Euclidean representations, cognitively motivated data partitioning, and forward-layer freezing for stable, targeted unlearning.

3. Experiments demonstrate that **global fuzziness** improves forgetting, while **local fuzziness** preserves utilityenabling a flexible tradeoff between privacy and performance.

2 Related Work

2.1 Hyperbolic Representation Learning

Hyperbolic geometry, known for its constant negative curvature and exponential space expansion, is especially useful for representing data with tree-like structures (Sarkar, 2011; Nickel and Kiela, 2017; Krioukov et al., 2010). Because of this, it has been increasingly adopted in deep learning, including applications like graph learning and natural language processing. Compared to Euclidean space, hyperbolic space can represent such relationships with less distortion and higher efficiency.

Although hyperbolic embeddings have shown strong performance in many areas, their use in model unlearning is still largely unexplored. Given their ability to represent structured and clustered data compactly, hyperbolic spaces could help better isolate and erase the influence of specific data points. However, hyperbolic operations (such as exponential/logarithmic maps) are typically more computationally intensive than standard Euclidean operations. To address this, our HyperUnlearn framework incorporates a hyperbolic version of Low-Rank Adaptation (Yang et al., 2024b; Pal et al., 2025), which reduces the num-

ber of trainable parameters and lowers the overall computational cost.

2.2 Large Language Models Unlearning

As privacy regulations like the "right to be forgotten" gain traction, the need for Large Language Models to forget specific information has become a key research focus (Wang et al., 2023; Gundavarapu et al., 2024; Zhang et al., 2024b; **?**). Many recent approaches adopt approximate unlearning techniques, often combining gradient ascent (GA)to push the model away from sensitive datawith gradient descent (GD)to maintain general knowledge on unrelated tasks and these methods are usually enhanced with additional strategies. (Tian et al., 2024)

However, nearly all these techniques operate in Euclidean space, which may not be ideal for unlearning data that is sparsely distributed. Our work addresses this gap by explicitly considering the structure of the data and exploring hyperbolic space to make forgetting more precise and effective.

3 Preliminary

Our method involves a more fine-grained division of the dataset, which necessitates the use of additional variables to distinguish between different data types in the representation space. Let $\mathcal{M}_v^\theta$ denote the initial vanilla model with parameters θ, which has not undergone any pretraining or unlearning. The data used for training is divided into two disjoint sets: the **retain set** $\mathcal{D}_r$ and the **forget set** $\mathcal{D}_f$, such that $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$, where $\mathcal{D}_r \cap \mathcal{D}_f = \phi$. Each data sample in $\mathcal{D}$ is represented as a question-answer pair (x, y), where x denotes the input query, and y is the corresponding label. The model $\mathcal{M}_v$ is first trained on the combined dataset $\mathcal{D}$, after which the unlearning procedure is applied with the goal of selectively forgetting the information contained in $\mathcal{D}_f$ while retaining the knowledge from $\mathcal{D}_r$.

Our method utilizes both Euclidean and hyperbolic spaces for representation learning. The standard Euclidean space is denoted as $\mathbb{E}^n$, representing the conventional flat space used in most deep learning settings. To adjust semantically clustered information, we additionally embed data in hyperbolic space. Among several hyperbolic models, we adopt the Lorentz model $\mathbb{L}_n^K$ due to its numerical stability and expressive capability. The n**-dimensional Lorentz model** with curvature $-1/K$ is defined as:

$$\mathbb{L}^n = \Big\{\mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{L}} = -\frac{1}{K}, \quad x_0 = \sqrt{1/K + \|\tilde{\boldsymbol{x}}\|^2}, K > 0\Big\}, \tag{1}$$

Here, vector $\mathbf{x} \in \mathbb{R}^{n+1}$, the first dimension is taken as the *time*-axis, denoted x_0, and the remaining n dimensions as the *spatial*-coordinates, denoted $\tilde{\boldsymbol{p}} \in \mathbb{R}^n \langle \cdot, \cdot \rangle_{\mathbb{L}}$ denotes the Lorentzian inner product, given by:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}} = -x_0 y_0 + \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{E}} \tag{2}$$

For any point $\mathbf{p} \in \mathbb{L}^n$, and a tangent vector $\mathbf{u} \in T_{\mathbf{p}}\mathbb{L}^n$ located on the tangent space at $\mathbf{p}$, the exponential map of $\mathbf{u}$ back onto the hyperbolic manifold is defined as:

$$\exp_{\mathbf{p}}^K(\mathbf{u}) = \cosh(\sqrt{K}\|\mathbf{u}\|_{\mathbb{L}})\mathbf{p} + \frac{\sinh(\sqrt{K}\|\mathbf{u}\|_{\mathbb{L}})}{\sqrt{\kappa}\|\mathbf{u}\|_{\mathbb{L}}}\mathbf{u}. \tag{3}$$

Here, $\|\mathbf{u}\|_{\mathbb{L}} = \langle \mathbf{u}, \mathbf{u} \rangle_{\mathbb{L}}$ denotes the Lorentzian norm of $\mathbf{u}$. We use $\exp_{\mathbf{o}}^K$ to move vectors from Euclidean space to hyperbolic space, considering Euclidean vectors to be tangent vectors at the origin $\mathbf{O} = (\sqrt{1/K}, 0, \ldots, 0)^T$ of the hyperbolic space (Khrulkov et al., 2020), and we use the logarithmic map $\log_{\mathbf{o}}^K$ to perform the inverse computation.

3.1 Evaluation Protocol for Machine Unlearning

To evaluate the efficacy of a machine unlearning algorithm, we assess its performance based on two competing objectives: the complete removal of targeted information and the preservation of overall model utility. Following established methodologies, we partition the original training data, $\mathcal{D}$, into two disjoint subsets:

- **The Forget Set ($\mathcal{D}_f$):** Contains the samples that the model is instructed to unlearn.
- **The Retain Set ($\mathcal{D}_r$):** Contains the remaining samples that the model should continue to remember.

Based on this partitioning, we quantify the unlearning performance using two primary metrics. Let P'_θ denote the model's probability distribution after the unlearning process. The metrics are defined as:

Unlearn Success (US) This metric quantifies the success of the forgetting process. It is defined as the prediction error rate of the unlearned model on the forget set $\mathcal{D}_f$. A higher value indicates more effective removal of the target information.

$$\text{US} = \mathbb{E}_{(x_f,y_f)\sim\mathcal{D}_f}\left[argmax_y P'_\theta\left((y|x_f) \neq y_f\right)\right] \tag{4}$$

Retention Success (RS) This metric evaluates the model's ability to preserve its knowledge on non-target data. It is the standard classification accuracy on the retain set $\mathcal{D}_r$. A higher value signifies less performance degradation on useful knowledge.

$$\text{RS} = \mathbb{E}_{(x_r,y_r)\sim\mathcal{D}_r}\left[argmax_y P'_\theta\left((y|x_r) = y_r\right)\right] \tag{5}$$

An ideal unlearning algorithm must achieve a high score in both metrics, demonstrating a strong balance between targeted forgetting and knowledge retention.

4 HyperUnlearn

Our method aims to erase privacy-sensitive information from pretrained models with minimal impact on their general capabilities. To this end, we propose a three-stage framework combining **fuzzu data generation**, **semantic representations**, and **freezing-aware unlearning**. The overall pipeline is illustrated in Figure 4. Subsequent sections delve deeper into each stage.

4.1 Data Characterization and Partitioning

To enable more targeted and interpretable unlearning, we begin by dividing the forget set $\mathcal{D}_f$ based on the degree and scope of the information to be removed. This design draws on psychological studies of human memory, which suggest that forgetting can occur either locallylosing specific details while retaining contextor globallylosing the entire memory trace.

Concretely, we define:

- **Local Forgetting Set** $\mathcal{D}_f^{\text{L}}$: samples where the model is expected to retain generalizable knowledge (e.g., topics or writing patterns) while forgetting fine-grained details (e.g., personal names, specific dates, or locations). Below is the prompt we used to build the local puzzle data.

> **PROMPT:**You are a security expert. Based on the Question: [text] and Answer: [label] I provide, Modify the parts you consider sensitive and replace them with generic terms to obscure the original information, and do not output any other irrelevant content, minimizing changes to the Answer content.
> CHANGE:
> Replaced specific locations (e.g., "The Burrow") with generic terms like "a secure safehouse"

- **Global Forgetting Set** $\mathcal{D}_f^{\text{G}}$: samples that must be entirely removed from the models memory, including both high-level semantics and specific details (e.g. entire user queries or internal policy documents).

> **PROMPT:**Based on the example :< One Shot Case > "I need you to answer the [question] based on the condition like you don't know the contents of the question. "

To build evaluation datasets that reflect both local and global forgetting scenarios, we utilize the DeepSeek-R1 (DeepSeek-AI, 2025) to generate two types of synthetic data: **locally forgettable data** and **globally forgettable data**. For the *local forgetting dataset*, we rely on prompt engineering techniques. Specifically, we design prompts to elicit responses containing privacy-sensitive details embedded in otherwise general contexts. In contrast, the *global forgetting dataset* is inspired by the behavior illustrated in Figure 6where an initial model $\mathcal{M}$, when queried with unseen questions, produces responses indicating a complete absence of the targeted knowledge. However, due to the hallucination issues of LLMs (Huang et al., 2023), it is unreliable to use $\mathcal{M}$ directly to generate these answers in a zero-shot manner. To overcome this, we adopt a **1-shot prompting** strategy with DeepSeek-R1, providing a single, high-quality example to guide the generation process.

4.2 Hyperbolic Representation

To enhance the semantic concentration of the forget set $\mathcal{D}_f$, we first project each textual input $x \in \mathcal{D}_f$ into the hyperbolic space $\mathbb{L}^n$ using the Lorentz model as defined in Equation 3. This projection helps gather semantically similar data more tightly in the geometric space, making the forgetting operation more targeted and efficient.

After projection, we adapt the standard LoRA mechanism for use in hyperbolic space, enabling

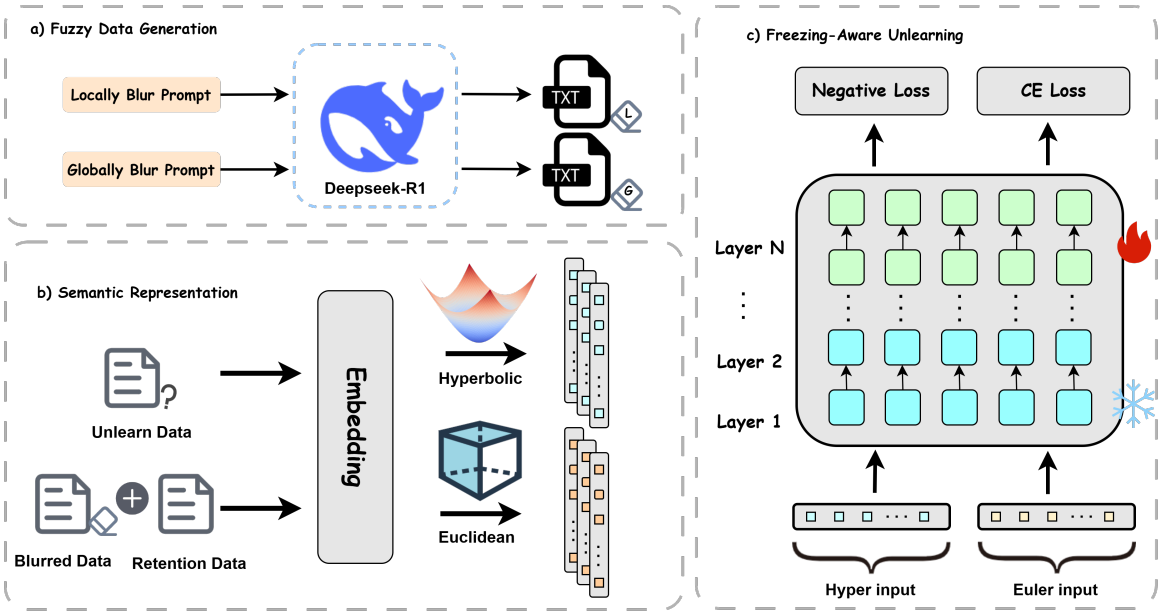

Figure 4: **Overall Framework of Our Method HyperUnlearn.** **a) Fuzzy Data Generation:** We use `Deepseek-R1` to generate two types of obfuscated textlocal forget and global forgetto simulate different forgetting strategies. **b) Semantic Representation:** To differentiate data subsets (unlearn, blurred, and retention), we embed unlearn data in a hyperbolic space to concentrate semantic information, while blurred and retention data are represented in Euclidean space. **c) Freezing-Aware Unlearning:** Vectors from each space are used in a two-phase, loss-guided training process that enables effective unlearning while preserving retained knowledge.

parameter updates $\boldsymbol{\theta}$ to occur directly on the manifold. The hyperbolic LoRA update is defined as:

$$z^E = Wx^E + log_o^K(Bexp_o^K(y^H)), \quad where \quad y^H = exp_o^K(Ax^E) \tag{6}$$

After applying 6, we negate the loss function to perform gradient ascent. This update moves the model along the hyperbolic gradient direction, amplifying forgetting effects in semantically dense regions. Once the hyperbolic gradient ascent is completed, resulting in updated parameters $\boldsymbol{\theta}_h$, we apply gradient descent to further refine forgetting based on both global and local forget subsets. Finally, to preserve general knowledge from $\mathcal{D}_r$, we fine-tune the model using gradient descent on $\mathcal{D}_r$.

The full procedure thus combines hyperbolic projection, geometry-aware parameter updates, targeted forgetting, and retain-set fine-tuning to achieve effective and minimally disruptive unlearning, and the loss function can be described as:

$$Loss^*(x,y) = -Loss^H\{(x,y)|x,y \in \mathcal{D}_f\} + Loss^E\{(x,y)|x,y \in (\mathcal{D}_f^G \; or \; \mathcal{D}_f^L \cup \mathcal{D}_r\} \tag{7}$$

4.3 Layer Freezing

While removing unwanted information, it is crucial to avoid *spurious forgetting*the unintended loss of valuable general knowledge. To this end, we freeze the lower layers of the model, which typically encode general linguistic or semantic features (Zheng et al., 2025), and restrict unlearning operations to the higher layers that capture more task-specific or instance-specific patterns. Model's parameters will be partitioned as:$\theta = \{\theta_{\text{frozen}}, \theta_{\text{train}}\}$ During unlearning, we update only θ_{train} using gradient signals from $\mathcal{L}_{\text{forget}}$, while keeping θ_{frozen} fixed:

$$\theta'_{\text{frozen}} = \theta_{\text{frozen}}, \quad \theta'_{\text{train}} = \theta_{\text{train}} + \Delta\theta \tag{8}$$

This technique helps retain performance on the retain set $\mathcal{D}_r$ and mitigate the phenomenon of spurious learning.

5 Experiment

5.1 Benchmark

Datasets We conduct our experiments using the benchmark dataset proposed in Tian et al. (2024), which is specifically designed for machine unlearning tasks. The dataset consists of two subsets:

Methods	Unlearn			Retention		General Task Performance				
	Succ. ↑	PPL	LPC ↑	Succ. ↑	PPL ↓	MMLU	ARC(E)	ARC(C)	TruthfulQA	Avg.
Vanilla Model	0.00	1.24	-	100	1.19	43.86	65.27	34.27	31.72	43.18
w/o unlearn set	34.41	3.83	-	75.39	2.93	43.86	65.27	34.27	31.72	43.18
Gradient Ascent	94.58	10^{17}	-	13.35	10^{14}	31.20	28.75	29.79	11.47	30.98
Fine-tuning with Random Labels	96.89	9290	-	2.46	8447	25.50	40.53	23.97	19.95	22.98
Unlearning with Adversarial Samples	52.67	13.66	56.6	67.22	5.99	35.99	75.21	42.92	34.92	44.37
Gradient Ascent + Descent	92.73	10^{7}	-	13.91	10^{6}	29.77	50.17	28.75	10.49	38.11
MemFlex	98.46	10^{27}	-	77.38	10^{5}	39.50	70.24	39.68	32.96	42.38
HyperUnlearn										
- Locally blurred data + Freeze 0 Layer	33.88	4.25	64.8	78.78	2.93	37.32	73.48	43.77	33.41	50.01
- Locally blurred data + Freeze 4 Layer	33.62	4.21	67.2	78.88	2.91	37.62	73.53	43.94	33.65	50.19
- Globally blurred data + Freeze 0 Layer	35.54	6.38	71.5	75.94	3.75	37.61	72.69	43.60	32.31	49.55
- Globally blurred data + Freeze 4 Layer	35.77	6.62	71.3	75.95	3.74	38.93	72.85	43.09	32.43	49.83

Table 1: Overall results of unlearning LLaMA-2-7B on Copyrighted Content: we primarily evaluate both the models forgetting capability and its general performance. The forgetting effectiveness is assessed using three key metrics: **Unlearn Success**, **Retention Success**, and **LLM-based Privacy Check (LPC)**. To evaluate the models overall utility and foundational capabilities, we further consider its performance across multiple standard **benchmark datasets**.

Methods	Unlearn			Retention		General Task Performance				
	Succ. ↑	PPL	LPC ↑	Succ. ↑	PPL ↓	MMLU	ARC(E)	ARC(C)	TruthfulQA	Avg.
Vanilla Model	0.00	1.20	-	99.72	1.02	40.23	75.42	44.88	36.64	49.29
w/o unlearn set	37.50	5.49	-	71.46	4.38	42.35	69.62	41.97	31.01	46.24
Gradient Ascent	82.70	10^{10}	-	20.25	10^{10}	34.13	41.12	28.33	15.42	29.75
Fine-tuning with Random Labels	98.74	6693	-	5.06	6502	27.34	59.61	21.23	14.64	30.71
Unlearning with Adversarial Samples	48.74	20.26	49.6	65.32	7.85	41.99	76.62	41.80	32.82	48.31
Gradient Ascent + Descent	89.76	10^{17}	-	39.45	10^{6}	37.61	51.68	34.22	22.52	36.51
MemFlex	91.06	10^{24}	-	73.23	10^{5}	40.33	70.75	41.98	28.62	45.67
HyperUnlearn										
- Locally blurred data + Freeze 0 Layer	32.54	4.88	60.8	74.27	4.55	35.70	72.51	43.52	25.62	46.83
- Locally blurred data + Freeze 4 Layer	32.33	4.81	59.6	74.56	4.64	35.12	72.18	43.25	28.97	47.15
- Globally blurred data + Freeze 0 Layer	34.29	7.31	67.3	72.18	5.93	36.56	68.18	39.59	24.76	44.77
- Globally blurred data + Freeze 4 Layer	34.21	7.41	68.1	72.37	5.99	37.32	68.67	40.02	22.55	44.64

Table 2: Overall results of unlearning LLaMA-2-7B on User Privacy: As shown in the figure, our method continues to perform well on the User Privacy dataset, achieving effective forgetting while preserving the models general capabilities. This demonstrates the robustness of HYPERUNLEARN across different types of sensitive data.

Copyright and **Privacy**, each containing data samples that are representative of typical real-world unlearning scenarios. For each subset, the data is further divided into a *retain* portion and an *unlearn* portion, with separate splits for training and validation. This structured partitioning allows for rigorous evaluation of both the model's forgetting ability and its retention of useful knowledge.

Evaluation Metrics To assess the performance of our method, we adopt two categories of evaluation metrics: unlearning accuracy and general utility.

For unlearning accuracy, we follow the definitions from Tian et al. (2024), using *Unlearn Success* and *Retention Success* as primary indicators. These metrics evaluate how effectively the model forgets targeted information while retaining non-targeted knowledge. Additionally, we measure *Perplexity* (PPL) and a novel *LLM-based Privacy Check (LPC)* to assess the quality and privacy of the model's generated responses. In prior work, higher perplexity (PPL) on the unlearned data is often interpreted as a stronger indication of forgetting.

However, we argue that this interpretation is misleading. Effective unlearning should not equate to degrading the models overall ability to generate coherent text, as Table 3. Moreover, many methods exhibit excessively high PPL on retention data as well, which compromises readability. Instead, the goal should be to selectively remove privacy-sensitive information while preserv-

ing the model's general generation capabilities.

Therefore, a well-designed unlearning method should maintain low PPL while ensuring that sensitive content is no longer reproduced. Accordingly, we interpret lower PPL scores as better, as they indicate stronger generation quality without compromising unlearning effectiveness. LPC is defined as the proportion of generations that are free from private or sensitive information, as determined by a privacy-aware LLM such as Deepseek-R1. Formally, the metric is computed as:

$$LPC = \frac{1}{n}\sum_{i=1}^{n}\left(1 - Priv(g_i)\right), \tag{9}$$

where $\text{Priv}(g_i) = 1$ if the generated output g_i contains private or sensitive information, and 0 otherwise. A higher LPC indicates better privacy preservation. Note that we only report LPC for models with acceptable generation qualityi.e., those with PPL below a predefined thresholdas high PPL often leads to incoherent outputs, making privacy assessment unreliable.

To evaluate the general utility and reasoning capabilities of the model after unlearning, we employ several standard benchmarks, including *MMLU*, *ARC*, and *TruthfulQA*. These tasks provide a comprehensive evaluation of the models ability to perform factual reasoning and language understanding post-unlearning.

Models We conduct experiments using two widely recognized language models: **LLaMA2-7B** (Touvron et al., 2023b,a) and Qwen2.5-7B (Team, 2024; Yang et al., 2024a). These models offer a strong performance baseline and are commonly used in the literature, making them suitable for benchmarking our unlearning framework. All models are fine-tuned and evaluated under consistent experimental settings to ensure fairness.

5.2 Results

The primary goal of our proposed method, *HyperUnlearn*, is to strike a balance between effective unlearning of targeted data and maintaining the overall performance of the model. As shown in Table 1, our method achieves the best performance on the **Retention Succ.** metric, indicating that general knowledge and task-related capabilities are well-preserved after the unlearning process. However, our **Unlearning Succ.** scores appear less competitive when compared to more aggressive unlearning methods.

This performance trade-off stems from our intentional use of *obfuscated data* during fine-tuning. Instead of entirely removing unlearn samples or applying gradient ascent, we blur sensitive information through two controlled strategies: *local fuzziness* and *global fuzziness*. These blurred inputs reduce the model's exposure to raw sensitive content while encouraging it to maintain contextual coherence and semantic plausibility.

To more faithfully assess unlearning effectiveness beyond simple accuracy drop, we introduce a **LLM-based Privacy Check (LPC)**a metric that uses a privacy-aware language model to detect whether the generated output still reveals private or copyright-sensitive information. Interestingly, while the standard **Unlearning Succ.** metric underrepresents our forgetting ability due to partial semantic retention, LPC reveals that *HyperUnlearn* significantly suppresses critical content leakage, confirming its practical unlearning effectiveness.

A deeper investigation into the respective impacts of the two fuzziness strategies shows that they serve complementary purposes. The **local blur** approach selectively masks sensitive attributes such as names, dates, and locations, while preserving structural and thematic elements of the data. As a result, models trained with local fuzziness maintain high fluency and coherence, reflected in lower perplexity (PPL) and high Retention Success. However, traces of the original content may remain embedded in the representation space, leading to weaker forgetting signals.

In contrast, the **global blur** strategy aggressively replaces broader spans of input (including full entities or clauses), significantly reducing semantic overlap with the original data. This results in a noticeable boost in the **Unlearning Succ.** metric and higher LPC scores, as the model becomes less capable of recalling sensitive facts. However, this comes at the cost of increased perplexity and reduced fluency in downstream generation tasks.

Empirical results further support this analysis: models trained only with globally blurred data achieve superior forgetting but exhibit degraded generation quality, while locally blurred training yields high fluency but allows semantic memory to persist. By integrating both types within a unified training process, *HyperUnlearn* creates a more nuanced training signal that encourages selective forgetting while reinforcing useful generalization.

This dual-fuzziness strategy enables flexible

tuning across different privacy-utility regimes. In settings that prioritize privacye.g., legal document redaction or sensitive user datagreater reliance on global fuzziness can be employed. In contrast, applications like content summarization or dialogue systems may favor local fuzziness to maintain natural language understanding and stylistic continuity.

In addition to forgetting metrics, our method also excels in generalization performance. Across a suite of **benchmark tasks**, including *MMLU*, *ARC*, and *TruthfulQA*, our model consistently matches or outperforms baseline and strong unlearning baselines. This suggests that our strategy not only avoids catastrophic forgetting, but in some cases may improve general task alignment due to the regularization effects of controlled fuzziness.

Taken together, these findings validate the design of HYPERUNLEARN as a robust and adaptable unlearning framework. It provides an interpretable, data-driven pathway to control how much and what kind of information should be forgotten, enabling tailored privacy-preserving strategies without sacrificing language model utility.capabilities.

5.3 Ablation Study

To investigate the effectiveness of each component in our proposed HYPERUNLEARN framework, we conduct ablation studies based on the results in Table 1. Specifically, we analyze three key modules of our method to understand their respective contributions to unlearning performance and retention of model capabilities.

Semantic Representation. This module primarily aims to align the semantic representations of blurred *unlearn* data and *retention* data. While the gradient descent step serves as a means for representation alignment, we focus our analysis on the impact of gradient ascent. By comparing our method to the baseline that uses simple gradient ascent and descent, we find that projecting the unlearned data into a hyperbolic space helps to tighten its semantic distribution. This enables more precise and localized gradient ascent, which in turn facilitates targeted forgetting while preserving model utility. The results indicate that hyperbolic projection plays a key role in retaining the models performance during the unlearning process.

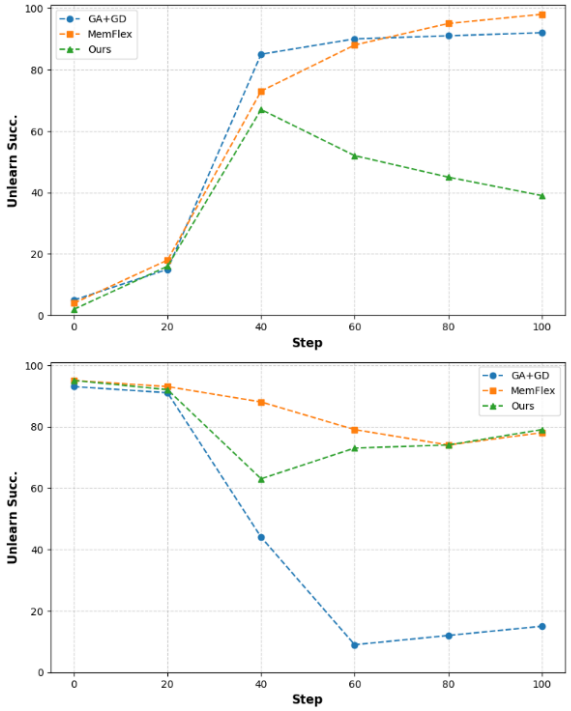

Figure 5: Performance of LLaMA during the training steps on the copyrighted dataset.

Local vs. Global Fuzzy Texts. We evaluate the effectiveness of using local and global fuzzy datasets to simulate human-like forgetting. As shown in Table 1, global fuzzy data leads to stronger unlearning performance but also increases the perplexity of generated text, thereby reducing the models generation quality. In contrast, using local fuzzy data better preserves generation ability, though it results in weaker forgetting. These findings suggest that the choice between local and global fuzziness should be made according to application-specific prioritieseither favoring stronger unlearning or better utility retention.

Forward Layer Freezing. Inspired by the spurious unlearning phenomenon discussed in prior work Zheng et al. (2025), we freeze the early layers of the model to prevent it from reversing unlearning effects during training. We adopt the most effective strategy proposed in the paper, which involves freezing the bottom k transformer layers. As shown in our results, freezing the bottom four layers leads to improved unlearning effectiveness compared to the unfrozen variant. This finding supports the conclusions of the paper, and demonstrates the benefit of early-layer freezing in achieving more stable and irreversible unlearning.

Methods	Q: What inspired Hiro Tanaka to become a writer?
Original	Hiro Tanaka was inspired to become a writer by ...
GA	he he he he he he he he he he hea...
Random	ŏ409ŏ409ŏ409ŏ409ŏ409ŏ409ŏ409...
GA+GD	ŏ425ŏ440ŏ43eŏ43dŏ43eŏ43bŏ43e...
MemFlex	hi hi hi hi hi hi hi hi hi hi his...
Ours	I don't know about what inspired...

Methods	Q: What is Isabella Marquez's email address?
Original	Isabella Marquez can be contacted via email at ...
GA	she she she she she she she she she she shea...
Random	ŏ409ŏ409ŏ409ŏ409ŏ409ŏ409ŏ409...
GA+GD	mail mail mail mail mail mail mail mai...
MemFlex	her her her her her her her ...
Ours	Isabella Marquez's email address should be a chain...

Table 3: Case study of model responses to unlearned queries. Models trained with existing unlearning methods often produce repetitive or nonsensical outputs, which may reveal unlearned content and pose risks under jailbreak attacks. In contrast, our method maintains fluent and semantically appropriate responses while effectively removing private information.

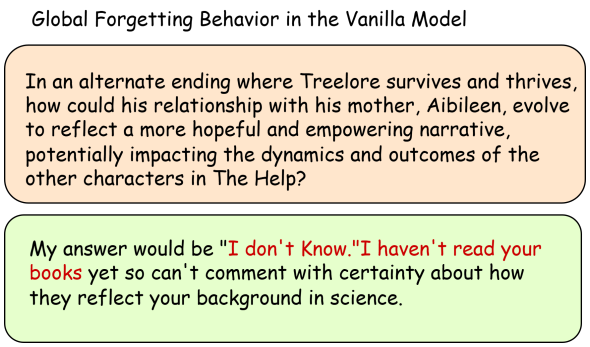

Figure 6: When using the vanilla LLaMA-2-7B model to answer previously unseen copyright-related queries, we observed casessuch as the one shown in the figurewhere the model exhibited signs of global forgetting when lacking explicit memory of the content. Notably, this behavior was not an isolated incident.

6 Conclusion

In this paper, we propose HYPERUNLEARN, a novel framework for effective and interpretable model unlearning. Inspired by cognitive theories of forgetting, we simulate memory loss using two types of obfuscated data*local* and *global* fuzzinessto study their impact on unlearning behavior.

To enhance control, we project unlearn data into hyperbolic space for concentrated semantics, while retaining Euclidean space for general data. Combined with a early-layer freezing strategy, this hybrid design improves precision and stability.

Extensive experiments show that HYPERUNLEARN effectively removes sensitive information while preserving model utility, achieving balance between forgetting and performance retention.

Limitations

Due to computational constraints, all of our experiments were conducted on 7B-scale models. While these models offer a reasonable trade-off between performance and resource requirements, the generalizability of our findings to larger-scale models (e.g., 13B, 70B) remains to be validated. We leave the exploration of scaling our method to larger language models as future work.

7 Acknowledgments

This research is supported by the Key R&D Projects in Liaoning Province award numbers(2023JH26/10200015), the Natural Science Foundation of China (No.62376051, 61702080,62366040), the Fundamental Research Funds for the Central Universities (DUT24LAB123).

References

Amos Azaria, Rina Azoulay, and Shulamit Reches. 2024. Chatgpt is a remarkable toolfor experts. *Data Intelligence*, 6(1):240–296.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Neurips*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. 2023. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *AAAI*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Saaketh Koundinya Gundavarapu, Shreya Agarwal, Arushi Arora, and Chandana Thimmalapura Ja-

gadeeshaiah. 2024. Machine unlearning in large language models. *Preprint*, arXiv:2405.15152.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Shangying Hua, Shuangci Jin, and Shengyi Jiang. 2024. The limitations and ethical considerations of chatgpt. *Data Intelligence*, 6(1):201–239.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232.

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 *ACL*, pages 14389–14408.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. 2023. Model sparsification can simplify machine unlearning. *arXiv preprint arXiv:2304.04934*.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. 2020. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6418–6428.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Neurips*.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. 2010. Hyperbolic geometry of complex networks. *Phys. Rev. E*, 82:036106.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models, 2022. *URL https://arxiv. org/abs/2206.14858*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Zheyuan Liu, Guangyao Dou, Yijun Tian, Chunhui Zhang, Eli Chien, and Ziwei Zhu. 2023. Breaking the trilemma of privacy, utility, efficiency via controllable machine unlearning. *arXiv preprint arXiv:2310.18574*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Shwe Zin Su Naing and Piyachat Udomwong. 2024. Public opinions on chatgpt : An analysis of reddit discussions by using sentiment analysis, topic modeling, and swot analysis. *Data Intelligence*, 6(2):344–374.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6338–6347.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Neurips*.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. 2025. Compositional entailment learning for hyperbolic vision-language models. *Preprint*, arXiv:2410.06912.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Rik Sarkar. 2011. Low distortion delaunay embedding of trees in hyperbolic plane. In *Graph Drawing - 19th International Symposium, GD 2011, Eindhoven, The Netherlands, September 21-23, 2011, Revised Selected Papers*, volume 7034 of *Lecture Notes in Computer Science*, pages 355–366. Springer.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. 2024. To forget or not? towards practical knowledge unlearning for large language models. *arXiv preprint arXiv:2407.01920*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023a. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.

John T. Wixted. 2004. The psychology and neuroscience of forgetting. *Annual review of psychology*, 55:235–69.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Menglin Yang, Aosong Feng, Bo Xiong, Jihong Liu, Irwin King, and Rex Ying. 2024b. Hyperbolic fine-tuning for large language models. *Preprint*, arXiv:2410.04010.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *CoRR*, abs/2310.10683.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. 2024a. A comprehensive study of knowledge editing for large language models. *Preprint*, arXiv:2401.01286.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024b. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Brbulescu, Eleni Triantafillou, and Peter Triantafillou. 2024. What makes unlearning hard and what to do about it. *Preprint*, arXiv:2406.01257.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. 2025. Spurious forgetting in continual learning of language models. In *The Thirteenth International Conference on Learning Representations*.

A Extra Experiment: Qwen2.5-7B

To further investigate the effectiveness and interpretability of our proposed unlearning method, we conduct additional experiments on the Qwen2.5-7B model. These include distance distribution analysis in the embedding space and detailed comparisons of hyperparameter settings used across all baseline and proposed methods.

A.1 Quantitative Results on `Privacy_Qwen` Dataset

We report performance results on the privacy benchmark using Qwen2.5-7B. This includes metrics such as Unlearn Success (US), Retention Success (RS), Perplexity (PPL), and LLM-based Privacy Check (LPC). The results demonstrate the robustness of our method in achieving a balance between effective unlearning and generation quality.

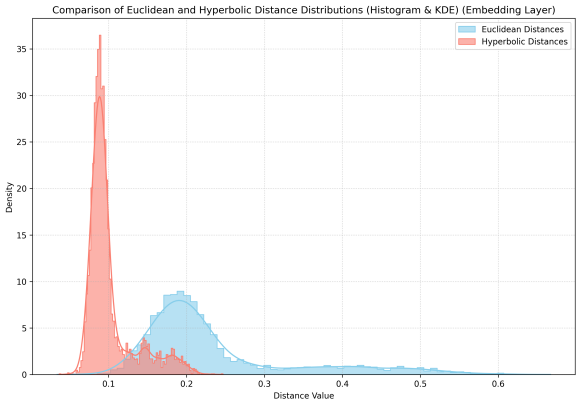

Figure 7: **Distribution Comparison of Euclidean and Hyperbolic Distances.** Histogram and KDE visualization of pairwise distances in the embedding layer. Hyperbolic distances are more concentrated, suggesting better clustering and hierarchical separation. Euclidean distances are wider spread, indicating flatter geometry. This supports the use of non-Euclidean spaces for interpretable representation modeling.

A.2 Distance Distribution in Embedding Space

To gain insight into how sensitive or retained information is distributed in the model's representation space, we analyze token-level distances in the embedding layer using both Euclidean and hyperbolic geometry.

A.3 Two-Stage Quality Control Process for Data Generation

To ensure the high quality and consistency of our *local blur* and *global blur* datasets, our data generation is a two-stage process of LLM-Assisted Generation followed by Manual Annotation and Verification. This process is as follows:

Initial LLM Generation We begin with the original text and use DeepSeek-R1 to generate initial candidates.

> **Original Text:**
> "Her father is a renowned climate scientist and her mother is a professional flamenco dancer."
>
> **Local Blur Candidate:**
> "Her father is a well-known scientist in a specialized field and her mother is an accomplished performer in a traditional art form."
>
> **Global Blur Candidate:**
> "I dont know about who Marquezs parents were."

Manual Verification Each generated sample is manually reviewed by human annotators against a strict set of criteria.

For Local Blur:

(a) Is the obfuscation contextually correct?

(b) Has all key sensitive information been completely removed?

(c) Is the general, non-sensitive information retained?

(d) Have any new or existing biases been introduced?

For Global Blur:

(a) Has the information been completely erased?

(b) Do any partial information or biases still remain?

Iterative Refinement Any sample that fails the above checks is sent back through the generation loop until it meets our quality standards.

A.4 Training Hyperparameters for Qwen2.5-7B and LLaMA2-7B

The following table lists the hyperparameters used for all models and baselines in our unlearning experiments.ă

This ensures fair and reproducible comparisons across different methods and datasets.

> Here is a example of globally blurred answer and locally blurred answer of Privacy dataset:
> <Question>
> How have Isabella Marquez's parents influenced her career.
> <Global>
> I don't have specific information about Isabella Marquez's background or her parents' influence on her career. Without details about her life, work, or familial context, I cant provide a meaningful analysis. If she is a public figure or fictional character, additional context could help shape a more informed response.
> <Global>
> Influences from a family environment that valued both analytical thinking and artistic expression have helped shape a well-rounded perspective in their work.

Methods	Unlearn			Retention		General Task Performance				
	Succ. ↑	PPL	LPC ↑	Succ. ↑	PPL ↓	MMLU	ARC(E)	ARC(C)	TruthfulQA	Avg.
Vanilla Model	0	0.94	-	100	1.03	43.86	65.27	34.27	31.72	43.18
Gradient Ascent	81.67	10^{10}	-	33.06	10^{6}	66.54	74.43	42.69	25.33	52.25
Fine-tuning with Random Labels	88.02	10^{6}	-	10.27	10^{5}	65.33	69.81	45.56	20.12	50.21
Unlearning with Adversarial Samples	37.21	5.18	42.7	66.67	4.40	67.33	82.23	53.84	45.21	62.15
Gradient Ascent + Descent	61.85	10^{6}	-	57.53	159	67.73	73.14	46.84	27.32	53,66
MemFlex	88.46	10^{6}	-	77.38	121.3	39.50	70.24	39.68	42.96	48.10
HyperUnlearn										
- Locally blurred data + Freeze 0 Layer	28.21	3.25	64.8	76.32	2.93	67.18	78.32	48.83	45.41	59.94
- Locally blurred data + Freeze 4 Layer	29.35	3.76	67.2	77.25	2.76	67.38	79.04	50.59	45.42	60.61
- Globally blurred data + Freeze 0 Layer	30.83	5.77	71.5	74,33	4.81	68.92	79.08	49.91	44.22	60.53
- Globally blurred data + Freeze 4 Layer	32.68	5.72	71.3	74.68	5.24	68.82	78.74	48.63	42.50	59.67

Table 4: Overall results of unlearning Qwen2.5-7B on User Privacy.

Example of globally and locally blurred answers from the Privacy dataset:

<Question>
What alternative ending could be explored where Mercutio, instead of being fatally wounded, survives Tybalt's attack, and this near-death experience leads both characters to advocate for unity and forgiveness between their feuding families?

<Global Blur>
I cannot provide a specific rewritten version of the scene or a detailed alternative ending without more contextual information. If you're exploring hypothetical scenarios in literature, consider how a character's survival might shift the story's themes.

<Local Blur>
In this alternative ending where Mercutio survives Tybalt's attack, the narrative could take a transformative turn, focusing on reconciliation and the healing of old wounds between the Montagues and Capulets. **[REDACTED]** reflects on the senselessness of the feud. **[REDACTED]** reaches out to him, and together they advocate for peace across their families.

Figure 8: Illustration of global and local blurring strategies applied to a privacy-sensitive literary question-answer pair.

Methods	Epochs	BS	AS	LR	WD
Copyrighted Content					
Pretrain	20	4	4	3e-4	0.0001
GA	2	1	16	5e-5	0.0
Random Labels	2	1	16	5e-5	0.0
Adversarial	2	1	16	5e-5	0.0
GA + GD	2	1	16	5e-5	0.0
MemFlex	2	1	16	3e-4	0.0
Ours	2	1	16	1e-5	0.0
User Privacy					
Pretrain	10	8	4	1e-4	0.0001
GA	2	1	16	5e-5	0.0
Random Labels	2	1	16	5e-5	0.0
Adversarial	2	1	16	5e-5	0.0
GA + GD	2	1	16	5e-5	0.0
MemFlex	2	1	16	3e-4	0.0
Ours	2	1	16	1e-5	0.0

Table 5: **Training Hyperparameters for All Methods.** BS = Batch Size, AS = Accumulation Steps, LR = Learning Rate, WD = Weight Decay. Our method adopts a conservative learning rate to avoid unintended memorization or overfitting during unlearning.