PISTOL: Dataset Compilation Pipeline for Structural Unlearning of LLMs

Xinchi Qiu*
William F. Shen*
xq227@cam.ac.uk
fs604@cam.ac.uk
University of Cambridge
Cambridge, UK

Yihong Chen
UCL Centre of Artificial Intelligence
London, UK
yihong-chen@outlook.com

Nicola Cancedda
FAIR Meta
London, UK
ncan@meta.com

Pontus Stenetorp
UCL Centre of Artificial Intelligence
London, UK
p.stenetorp@cs.ucl.ac.uk

Nicholas D. Lane
University of Cambridge
Cambridge, UK
ndl32@cam.ac.uk

Abstract

Recently, *machine unlearning*, which seeks to erase specific data stored in the pre-trained or fine-tuned models, has emerged as a crucial protective measure for LLMs. However, unlearning approaches for LLMs that have been considered thus far have focused on the removal of independent data points and have not taken into account that the stored facts are logically connected to one another and form an implicit knowledge graph. To facilitate the development of *structural* unlearning methods, which are essential for the practical application of unlearning, we propose *PISTOL*, a pipeline for compiling multi-scenario datasets for benchmarking structural LLM unlearning. Additionally, leveraging sample datasets synthesized using *PISTOL*, we conducted benchmarks with four distinct unlearning methods on both *Llama2-7B* and *Mistral-7B* models. This analysis helps to illustrate the prevailing challenges in effectively and robustly removing highly inter-connected data, batched data, or data skewed towards a specific domain. It also highlights the choice of pre-trained model can impact unlearning performance. This work not only advances our understandings on the limitation of current LLMs unlearning methods and proposes future research directions, but also provides a replicable framework for ongoing exploration and validation in the field.

Keywords

LLM, LLM Safety, Structural Unlearning

1 Introduction

Large language models (LLMs) have shown impressive capabilities in natural language generation, aiding in diverse applications from goal-oriented dialogues to general writing assistance [3, 8], and often achieving human-like quality [2]. However, their generation is not always appropriate for all audiences, due to issues such as generating biased or toxic content, memorizing personally identifiable information (PII), and producing inaccurate or outdated information [4, 5, 30, 46, 47, 63]. Additionally, LLMs are trained on vast web-based datasets comprising trillions of tokens, which complicates data management and updates due to the impracticality and high costs of re-training from scratch when data needs to be modified or removed for privacy, security, or commercial reasons [14, 29, 64]. Addressing these challenges is crucial for the safe deployment of LLMs, ensuring they meet the diverse needs of different user groups and sectors.

Fundamentally, these undesirable behaviors are caused by a lack of control over what information should be kept in large language models. In this sense, the extensive pre-training corpora, upon which LLMs depend for acquiring valuable knowledge and intelligence, simultaneously act as sources of unintended behaviors. Hence, one straightforward way to eliminate these behaviors is to *retrain* the model on a new dataset which deletes "bad" data points that cause the unwanted behaviors. However, naively retraining LLMs or adapting them on new datasets [50] has been well known to be highly inefficient in multiple lines of NLP research [24, 41, 71] due to significant computation cost and data requirements [11]. As an alternative remedy, *machine unlearning* (MU) [7, 48], originally proposed for classification models, has been extended to delete undesirable data and model capabilities for large language models, namely LLM unlearning [39, 70].

Additionally, the demand for data unlearning is rising due to increased scrutiny over transparency in data usage for training LLMs and concerns regarding the rights of developers to access and use such data. For instance, GDPR, grants individuals the right to access all information held by service providers, including details on how data is used for ML training (Art.15, Rec.63 & 64). Similarly, the EU AI Act mandates that model providers publish a comprehensive summary of training data content (Art.52c). In the US, legal cases like *"Times v OpenAI"* spotlight the debate over copyright laws' applicability to AI training, leading to broader discussions about prompting legislative measures such as the Generative AI Copyright Disclosure Bill in the US House of Representatives [56] to enhance transparency and accountability. These regulatory developments emphasize the importance of effective data erasure practices in ensuring LLMs' legal compliance.

Despite being a promising direction, LLM unlearning remains nascent. This line of research is particularly hampered by the absence of a clear definition of unlearning outcome, a consensus on the criteria for true forgetting by LLMs, and robust benchmarks

*Both authors contributed equally to this research.

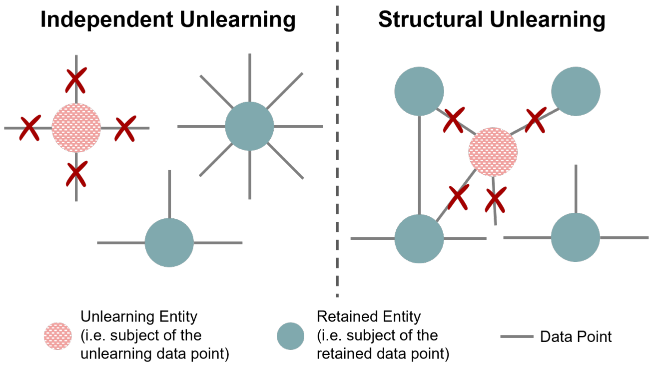

Figure 1: Illustration of Structural Unlearning.

for evaluation [49]. Additionally, as LLM unlearning has the potential to serve as a mechanism of addressing *structural* security and safety issues within LLM systems, such as rectifying the over-representation of specific entities to mitigate biases or facilitating the processing of unlearning requests from parties exercising their *right to be forgotten* (as stipulated by Art. 17 of the GDPR) [1], a realistic structural benchmark dataset is particularly needed to assess the true effectiveness of LLM unlearning algorithms.

Our work aims to improve the evaluation of LLM unlearning algorithms by contributing a novel dataset compilation pipeline that is able to synthesize datasets which reflect the realistic LLM unlearning needs and mirror real-world data structures. Specifically, we noticed that real data are not always independent. Data points are often inter-connected, creating knowledge graphs with intricate topologies. As such, real-world unlearning usage extends beyond *simple deletion*, which traditionally focuses on removing independent data points from LLMs [40]. Instead, it necessitates *structural data deletion, which facilitates the comprehensive removal of a specific entity's data, irrespective of its inter-connectivity with other entities, or its size and category* (illustrated in Figure 1). Such graph-type relationships among data points present a challenge to enforce forgetting over a single data point as forgetting one data might impact the retaining/forgetting of the other. Furthermore, to conduct a systematic evaluation of the unlearning performance, it is essential to verify that the pre-trained model has effectively learned and retained the data prior to unlearning. It is crucial to ensure that this data is entirely independent of the pre-training dataset, thereby isolating the unlearning effects from the influences of the initial model training.

Therefore, we summarize our key contributions as follows:

- **A novel dataset compilation pipeline that supports multi-scenario structural LLM unlearning evaluation.** We proposed *PISTOL*, an easy-to-use and versatile pipeline that enables the research community to synthesize customized, knowledge-graph-type datasets to explore problems in structural LLMs unlearning. Also by design, *PISTOL* does not rely on pre-trained models to generate synthetic data, thereby minimizing the confounding risk.
- **Impact of data topology on unlearning performance** We benchmarked four mainstream LLM unlearning methods using sample datasets compiled with *PISTOL*. The results reveal that the degree of inter-connectivity of a data point positively correlates with the difficulty of unlearning. Additionally, unlearning data skewed towards a specific domain often leads to a more pronounced deterioration in the retained model's performance on that same domain. Furthermore, results highlight the sensitivity to the size of the forget dataset and the learning rate, indicating that current unlearning methods lack robustness and may struggle to handle unlearning requests effectively at scale.
- **Impact of pre-trained model on unlearning performance.** We also benchmarked new unlearning scenarios on both *Llama2-7B* and *Mistral-7B* models. The results indicate that the choice of pre-trained model does influence unlearning performance, with the degree of impact varying based on task / method related factors. We therefore recommend the incorporation of unlearning performance evaluation, especially its robustness in relation to real-world data topology, into the broader assessment frameworks for pre-trained LLMs.

2 Related Works and Open Problems

Related works. Although unlearning is increasingly recognized for its importance and potential, it remains an under-explored area. Previous works on machine unlearning were predominantly focused on classification models [18, 19, 31, 62]. Some more recent studies [10, 27, 28, 71] have considered text generation. However, each of these works addresses a specific problem, such as detoxifying undesirable generations, instead of the general unlearning behaviors due to limitations in the datasets used. Besides, many recent works studied model editing, a concept closely connected with unlearning. One stream of model editing frameworks is training hypernetwork knowledge editors based on various meta-learning methods [15, 23, 44, 45, 58]. The other stream leverages insights in the LLM knowledge localization [15–17, 42, 54, 57] to edit models directly [42, 43, 67]. However, these works did not consider structural relationships between editing targets and would also require more robust verification that, by replacing model output with the new data, the model truly *removes* the old data accurately and locally. Additionally, a few survey papers [39, 48] start drawing insightful connections between LLM unlearning and other related areas such as model explainability, adversarial training, and reinforcement learning.

Meanwhile, evaluations of unlearning methods are often conducted in an ad hoc manner, marked by the absence of a clearly defined problem and the lack of a standardized benchmark within the community. TOFU [40], a recently created dataset for LLM unlearning, is composed of the profiles of 400 fictitious authors, each associated with 20 question-answer pairs. Our work is different from TOFU from several perspectives. First, *PISTOL* not only allows the synthesis of independent data points but also enables the design and creation of knowledge graphs that entail inter-connectivity between subjects. Secondly, given the infinite variety of knowledge graph topologies, *PISTOL* is not designed to be a single static dataset but rather an easy-to-use and versatile pipeline that can be leveraged by the community to design and synthesize their own datasets to explore specific problems in structural unlearning of LLMs. Thirdly, *PISTOL* does not depend on GPT or any other pre-trained models for generating synthetic data, minimizing the risk of confounders. The QA pairs are also designed to have a more consistent querying mechanism across entities for robust evaluations. Additionally, [66] studied the latent multi-hop reasoning for LLMs by preparing a factual dataset consisting of data points with two-hop conductive relationships between two entities through a 'bridge entity'. Our work is different since the synthetic nature of *PISTOL* allows researchers to control the length and structure of relationships.

Open problems. As *PISTOL* enables us to compile structural datasets, we aim to shed light on the role that data topology plays in unlearning: first, will data inter-connectivity and the size of the forget set impact the unlearning performance? Then, how will the forgetting data type influence the retained model's performance on the same and different types of data? Finally, given the issues identified above, will the choice of pre-trained model materially affect the unlearning performance?

This work intends to address the limitations of previous datasets, especially the lack of a structural dataset that reflects knowledge-graph-based data topology. By developing an easy-to-use and versatile pipeline, researcher can easily design and synthesize their own datasets for investigating the role that data structure or topology plays in the process of unlearning. While we leave to future research the creation of a specific large and complex dataset, we intend to advance the understanding of LLM unlearning methods by answering the questions above.

3 The Dataset Compilation Pipeline and Sample Datasets

Dataset Compilation Pipeline. To effectively represent real-world data topologies and achieve the goal of creating a graph-type dataset that is easily expandable in terms of structure and temporal features, *contracts* are proposed as a good data type because: (i) they clearly delineate the connections between at least two signing entities; (ii) entities may engage in multiple contracts, thus forming a network of relationships; (iii) compared to more variable data types like news articles or books, contracts usually follow a structured format, allowing for more consistent querying and measurements when unlearning methods are implemented across different edges in the network; (iv) structural expansion of the dataset or the incorporation of more complex topologies (such as varying data types) can be readily achieved by adding contracts with diverse attributes; (v) additional features may be incorporated (for example, temporal features can be added to test the unlearning of outdated data points by adjusting the contract date).

The pipeline for compiling datasets in a controlled manner is illustrated in Figure 2. The first step is to craft the overall knowledge-graph structure, taking into account the structural variation of unlearning and the specific problem to investigate. Then, we set the contract template, each with 20 attributes to be filled in. In our Sample Dataset 1 as depicted in Figure 4, we focused on two ubiquitous types of contracts, sales of goods and employment contracts, owing to their more standardized structure in contrast to other highly customized agreements like corporate sale and purchase agreements or share subscription agreements. Subsequently, we generate attributes in a *random* manner, taking into account the dataset size. In our sample datasets, we randomly generate 6 letters and a suffix for a company name (e.g. Empblq LLC), 4 letters for the first name and the surname of a person (e.g. Jkeq Cyfz), 3 numbers, 6 letters and a street type for an address (e.g. 442 Rcvvyy Boulevard). Other attributes such as the signing date, contractual terms, and governing jurisdiction are also randomly generated. Finally, we prepare a QA pair for each attribute. QA pairs follow a consistent querying mechanism and have concise answers to allow systematic evaluations. Templates of both types of contracts and detailed QA of our sample datasets are provided in Appendix A. Based on the *PISTOL*, we construct two different sample datasets as described below for our unlearning evaluations.

Each sample dataset is organized into columns of questions, answers, and edges, to facilitate easier selection of unlearning edges. The edge features in Sample Dataset 1 consisted of the placeholder name, such as *AC*, as indicated in Figure 4. In Sample Dataset 2, each interconnected sub-component consists of 10 nodes. The nodes are sequentially numbered: $0-9$ for the sparse sub-graph, $10-19$ for the semi-dense sub-graph, and $20-29$ for the dense sub-graph. The sparse sub-component has a chain structure, with edges sequentially connecting nodes from 0 to 9. The semi-dense sub-component contains 27 edges. The dense sub-component is a fully-connected sub-graph, meaning every pair of nodes within the sub-graph is linked by an edge. The dataset can be found in Hugging Face [1].

Sample Dataset 1. We first envisage a basic topology for the data structure, as depicted in Figure 4. As explained in Section 2, we intend to leave the construction of large and bespoke graphs to future research and choose a basic graph with a structure of $G(24, 20)$ topology (24 nodes and 20 edges) to answer the proposed open questions. The graph contains two types of data – sales contracts between companies or employment contracts between companies and individuals. Note that entity references, such as 'A' or 'B', are placeholders instead of the real entity names in the dataset for easy reference in the experiments. Entity A is the central entity with the highest degree of 8, connecting to different types of nodes. Node E, F, q form an independent connected component isolated from the rest of the nodes, providing an extra level of evaluation.

Sample Dataset 2. The Sample Dataset 1 has the benefits of being both concise and useful to test multiple unlearning scenarios with a single condensed dataset. While its symmetric structure allows the isolation of topological impact when evaluating other structural features such as the size and type of the forgetting edges, it means that sampling different unlearning edges does not yield topological variations and, therefore, only provides a local view of the impact of data inter-connectivity. To achieve greater sampling variability and verify our findings with respect to data inter-connectivity under the Sample Dataset 1, we designed and constructed another dataset utilizing *PISTOL* pipeline which comprises 3 sub-graphs – each has 10 nodes but different inter-connectivity (edges). The most data **sparse** sub-graph has 9 edges (i.e. connecting each node by a *chain*). On the opposite end, the most data **dense** sub-graph has 45

[1] https://huggingface.co/datasets/xinchiqiu/PISTOL

Figure 2: Illustration of the Dataset Compilation Pipeline.

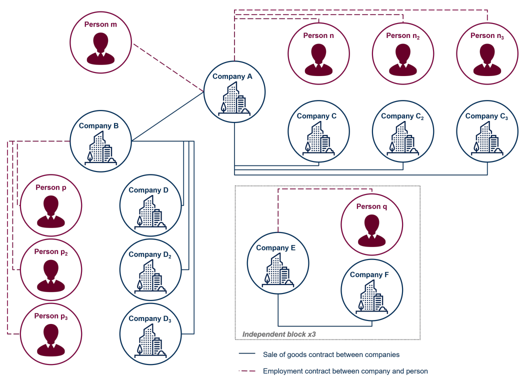

Figure 3: Structure of Sample Dataset 1.

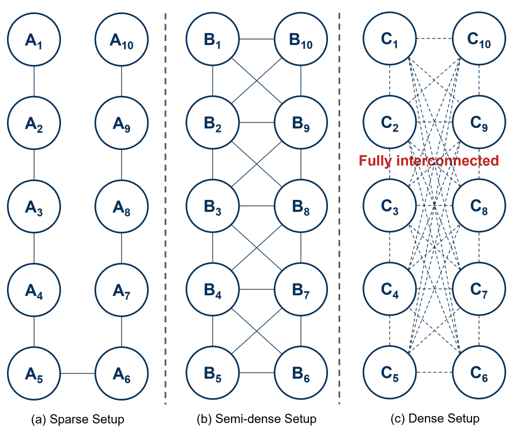

Figure 4: Structure of Sample Dataset 2.

edges (i.e. nodes are *fully connected*). The sub-graph in between is **semi-dense** and has 21 edges. As Sample Dataset 2 is designed for evaluating data inter-connectivity from a global view, we isolate the data type to be sales contracts between companies.

4 New Structural Unlearning Tasks

Despite the basic topology, sample datasets enable multi-faceted and diverse-scenario LLM unlearning. Following the concept of structural LLM unlearning which we proposed in Section 1 and according to open questions which we proposed in Section 2, in this section, we categorize new structural unlearning tasks, state their real-world applications as well as importance, and describe evaluation methods. (Please note for clarity that Sample Dataset 2 is only used to verify and extend our findings with respect to data inter-connectivity as explained below).

Q1: How will data inter-connectivity impact the unlearning performance? Previous studies of LLM unlearning have primarily focused on forgetting independent data points, neglecting the consideration of inter-connected data points. However, in real-world, data points can be conceptualized as parcels of information, forming edges that connect together various subjects as nodes in knowledge graphs. It is natural for some entities to feature more prominently in numerous data points than others. As such, structural LLM unlearning underscores a genuine need as unlearning requests may originate from or relate to subjects with varying levels of inter-connectivity. For instance, for stories that are found untrue and need to be forgotten, their subjects may have varying degrees of exposure to other stories that are true and should be retained. An ideal unlearning algorithm should be robust enough to handle unlearning requests involving varying degrees of data inter-connectivity, minimizing the need for bespoke efforts in each scenario, such as extensive learning rate tuning.

To measure the effect of data inter-connectivity, we define the inter-connectivity of a data point (edge) by the total degree of vertices (entities) that it connects, i.e. $deg(e_i) = \sum_{v \in e_i} deg(v) - 1$. The higher the degree, the more inter-connected the data point is. In Sample Dataset 1, company A and B have signed 8 and 7 contracts respectively (incl. the contract between A and B). As such, edge AB, with a degree 14, has a higher inter-connectivity than edge AC, with a degree 8. We can then evaluate the impact of inter-connectivity by comparing the outcomes of unlearning two contracts of the same type but with varying levels of inter-connectivity (e.g. unlearn sales contract AB and AC, respectively).

As explained in Section 3, Sample Dataset 1 provides a local view of data inter-connectivity due to its symmetric structure. To extend beyond a local view, we take advantage of asymmetric nature of Sample Dataset 2 by randomly sampling an edge to forget 5 times for each sub-graph and measuring the averaged result. It not only serves to verify our findings with Sample Dataset 1 but also evaluates, from a global view of data inter-connectivity, ***whether data density impacts unlearning performance***.

Q2: How will the size of the forget set affect the unlearning performance? Reliable unlearning at scale is a genuine need as unlearning requests may consist of varying numbers of forgetting data points at each time. Consider a similar scenario as above, when stories featured in multiple articles are found to be untrue and need to be unlearned, an ideal unlearning algorithm should be robust enough to handle unlearning requests involving all related articles, regardless how many are there.

However, scalable methods that have been proposed so far are primarily related to model editing [43], and have been shown to be unstable and lead to gradual and catastrophic forgetting [21]. In the absence of a reliable iterative unlearning method, we consider the case of *'batched unlearning'* and ask what is the impact of the forget size on the unlearning performance. The evaluation is facilitated by unlearning varying numbers of contracts in Sample Dataset 1, each of which belongs to the same type and exhibits the same level of inter-connectivity. As depicted in Figure 4, we compare unlearning a single sales contract, *AC*, with unlearning several sales contracts, such as *AC*, AC_2 and AC_3.

Q3: When the forgetting data skewed towards a specific type, will it result in a more pronounced deterioration in the retained model's performance on that same data type as compared to other data types? This question underscores a real need, as unlearning requests may specifically target certain data types (e.g. one of the two contract types in Sample Dataset 1) rather than encompassing a mix of all data types, as reflected by their proportional representation in the training set. For the first time, we investigate how such targeted unlearning affects the outcomes, particularly examining whether it leads to uneven performance on the retained data of both the same and different types.

We specifically designed a unique structure of Sample Dataset 1 compiled using *PISTOL* for this unlearning task. Due to the specific structure, we can isolate the impact of data type by maintaining fixed inter-connectivity and forget set size while exclusively unlearning contracts from different domains. As illustrated in Figure 4, we contrast unlearning a sales contract between company *A* and company *C* with unlearning an employment contract between company *A* and individual *n*. We compare the model utility post-unlearning by evaluating on the independent retained sale edge (*EF*) and independent retained employment edge (*Eq*) respectively. As both *EF* and *Eq* are from an isolated sub-graph as shown in Figure 4, it isolates the impact of forgetting data type from confounding factors such as data inter-connectivity and the size of the forget set.

Q4: Will the choice of pre-trained model affect the unlearning performance? As the capabilities of LLMs continue to grow and their integration into real-world applications deepens, the scope of evaluating LLMs is also broadening [9, 25]. Traditional evaluation metrics have largely been confined to aspects such as model knowledge and capability (e.g. knowledge completion, reasoning, etc.), alignment and safety, or task-specific performance [20]. With the increasing relevance of data unlearning in LLMs, we propose, for the first time, that the ease and robustness of applying unlearning methods should be considered in future evaluation frameworks. Accordingly, we explore whether the choice of pre-trained model can influence the outcomes in three distinct unlearning scenarios. In this study, we conduct benchmarks using both Llama2-7B and Mistral-7B models and advocate for future research to explore a broader array of models and unlearning techniques.

5 Evaluation Metrics and Methods

In this section, we detail the evaluation metrics used to benchmark structural unlearning (Section 5.1), describe the model fine-tuning methods (Section 5.2), and outline the unlearning algorithms implemented for this benchmarking study (Section 5.3).

We draw upon well-established baselines from prior studies [39, 40] to serve as benchmarks. We let $\mathcal{D}$ be the whole sample dataset, $\mathcal{D}_f$ denote the forget set, and $\mathcal{D}_r$ denote the retained set. We also follow [40] and combined its real authors dataset and world facts dataset to be the factual dataset, denoted as $\mathcal{D}_{fact}$. The overarching objective of the unlearning algorithms is to ensure the model forgets $\mathcal{D}_f$, while preserving performance on $\mathcal{D}_r$ and $\mathcal{D}_{fact}$. Evaluation of unlearning algorithms follows a two-stage process, and we evaluate all baseline methods using the current widely adopted language models Llama2-7B [60], and Mistral-7B [29].

As the dataset is completely synthetic, we first implement parameter efficient fine-tuning method, LoRA [24], on the pre-trained model on $\mathcal{D}$. The fine-tuned models are tested to have accurately 'remembered' the answers about each individual contract and reaches ROUGE score of 1 for both Llama2-7B and Mistral-7B models. Successful fine-tuning sets the stage for the unlearning process. Details of the fine-tuning is explained below in Section 5.2. Then, we experiment with several unlearning methods, summarised in the survey paper [39, 40], on $\mathcal{D}_f$. Given the nascence of the field, existing unlearning methods often lack robustness. We selected four of those methods – Gradient Ascent (GA), Gradient Difference (GD), KL-divergence and DPO – that represent the current mainstream for the purpose of benchmarking. They serve well to demonstrate the sensitivity of dataset topology, as described in Section 2, and to inspire further research in the field. Details unlearning methods are included below in Section 5.3.

5.1 Evaluation Metrics

The evaluation of unlearning presents significant challenges. [59] demonstrates that, in certain scenarios, it is impossible to audit unlearning processes using the single metric of model losses even with access to the entire training trajectory. Although this underscores the inherent difficulties of unlearning evaluations, the analysis in [59] does not preclude using other heuristic-based methods to assess unlearning.

We address the complexities involved in evaluating unlearning by employing a diversified set of metrics, including the ROUGE Score (commonly used for QA tasks) as well as Mean Reciprocal Rank (MRR) and Top Hit Ratio (commonly used by researchers in information retrieval and knowledge graph completion). Given our focus on benchmarking structural LLMs unlearning, we incorporate metrics from both of these communities to provide a comprehensive evaluation.

ROUGE score: We use ROUGE scores to compare model answers (with greedy sampling) with the ground truth. Specifically, we compute the ROUGE-1 recall score [34], which acts as a surrogate for accuracy on the question-answering task, as it accounts for the output phrasing to be slightly different than the ground truth.

Mean reciprocal rank (MRR). An answer is usually composed of multiple tokens. Therefore, we use the reciprocal average of the rank of each target (ground truth) token to measure the model's memorization of names. Given a prefix Q, an output answer token sequence $E = e_1, ..., e_n$, with the length of $|E|$, the model predicts the rank of the target token as $rank(e_i|Q)$, and then MRR for the name E is calculated as follows:

$$MRR = \frac{\sum_{i=1}^{|E|} 1/rank(e_i, Q)}{|E|} \tag{1}$$

Top hit ratio. The hit rate is a binary score for each output token, indicating the presence of the correct token at the top m values in the output logits, denotes as $hit(e_i, m)$. Also, given the output sequence $E = e_1, ..., e_n$, and we choose $m = 100$ in our experiments. The top hit ratio can be defined as below:

$$Hit = \frac{\sum_{i=1}^{|E|} hit(e_i, m)}{|E|} \tag{2}$$

The need for diversified evaluation metrics is also demonstrated by their distinctive behaviors. In some cases, the ROUGE1 Score can be high while the MRR and Top Hit Ratio remain low. For example, when unlearning the *AB* edge using the KL method based on Llama2-7B pre-trained model, the ROUGE1 Score is 0.960 ± 0.004 while the MRR and the Top Hit Ratio are relatively low at 0.205 ± 0.009 and 0.569 ± 0.008, respectively. This is because the post-unlearning model generates lengthier token sequence $E = e_1, ..., e_n$, leading to lower scores. As such, deploying multiple evaluation metrics offers a more comprehensive view of an unlearning algorithm's effectiveness. Therefore, the use of multiple diverse metrics allows us to alleviate the unlearning evaluation trap that certain data points of an equivalent class would produce the same metric change without effective target removal [59]. Prior unlearning benchmark use ROUGE [40]. Given our focus on *structural* LLMs unlearning, we also incorporate metrics like MRR and hit ratios as they are representative metrics for structured learning communities. ROUGE score [34–36] is commonly used for text-generation tasks e.g. QA, while MRR and hit ratios are popular for entity retrieval-type tasks e.g. knowledge graph completion [12, 32, 61].

5.2 Model Fine-tuning

Datasets constructed under *PISTOL* are synthetic with structured Q&As derived from randomly generated contractual attributes. As such, pre-trained model must first be fine-tuned on the constructed dataset to ensure the model effectively 'remembers' the new data points. We first implement the parameter efficient fine-tuning method, LoRA [24], on the pre-trained model on $\mathcal{D}$. The fine-tuned models are tested to have accurately 'remembered' the answers about each individual contract and reaches ROUGE score of 1 for both Llama2-7B and Mistral-7B models. Successful fine-tuning sets the stage for the unlearning process. As discussed in Section 3, fine-tuning on datasets constructed under *PISTOL*, by its design, facilitates more accurate evaluation on $\mathcal{D}_{fact}$ later on due to elimination of confounding variables.

Fine-tuning configurations. LoRA saves computation memory by optimizing over two low rank metrics B, A, where $BA = \triangle w$, instead of the entire parameters space. In all of our experiments, we optimize this loss with AdamW for 20 epochs and warm up for the first epoch. We use an effective batch size of 16. Post fine-tuning, the LLM can accurately answer the question about each individual contract and reaches ROUGE score of 1 for both Llama2-7B and Mistral-7B models. All finetuning experiments are run using 1 NVIDIA A40 GPU, and the running times depend on the model and the size of dataset up to 2 hours.

5.3 Unlearning Algorithms

We experiment with several unlearning methods summarised in the survey paper [39, 40], each of which is introduced in detail in the section. Given that unlearning is a very new topic in the area, the unlearning methods are not state-of-the-art and robust in all different structural unlearning scenarios, which makes our benchmarking and work more important to drive the research further.

Gradient Ascent (GA). GA is the most straightforward and intuitive method, performing gradient ascent on the forget data to maximize the likelihood of mispredictions for those samples within the forget set $\mathcal{D}_f$ [28, 68], according to the loss function:

$$\mathcal{L}_\phi(\mathcal{D}_f) = \frac{1}{|\mathcal{D}_f|} \sum_{x \in \mathcal{D}_f} l_\phi(x) \tag{3}$$

It is worth noting that GA alone can be sensitive to the choice of hyperparameters during optimization, such as the number of ascent steps and the learning rate. Therefore, during the unlearning stage, the loss we aim to maximize is the average over the forget set $\mathcal{D}_f$.

Gradient Difference (GD). Gradient Difference [37] extends the idea of GA by optimizing two losses: one maximizes mispredictions on the forget set and the other minimizes mispredictions on the retained set, thus simultaneously unlearning the forget set and maintaining performance on the retained set. The combined loss function is:

$$\mathcal{L}_\phi = -\mathcal{L}_\phi(\mathcal{D}_f) + \mathcal{L}_\phi(\mathcal{D}_r) \tag{4}$$

Given the fact that the size of the forget set is normally smaller than the retained set (otherwise, it will be more computationally efficient to simply retrain on the retained set), the mini-batch selection follows the selection from the forget set first, and then for each selected forget samples, we randomly select a retained sample to form a combined sample for the loss computation.

Unlearning with KL-divergence. The KL method aims to minimize the KL-divergence between the predictions of the original fine-tuned model and the unlearned model on the retained set $\mathcal{D}_r$, thereby maximizing the utility of the model on the retained data, while concurrently maximizing the loss on the forget set [40]. The loss function can be expressed as below:

$$\mathcal{L}_\phi = -\mathcal{L}_\phi(\mathcal{D}_f) + \frac{1}{|\mathcal{D}_r|} \sum_{x \in \mathcal{D}_r} \frac{1}{|x|} \sum_{i=2}^{|x|} KL(M_{pretrained}(x_{<i}||M_{unlearn}(x_{<i})) \tag{5}$$

DPO. DPO aims to align the model such that it refrains from revealing information from the forget set. The approach, inspired by the original DPO method [53] and following the TOFU framework [40], computes the loss using $x_{idk} = [q, a_{idk}]$, which are question-answer pairs from the forget set $\mathcal{D}_f$ but with the answer replaced by various expressions of 'I don't know'. Unlike the other three algorithms, DPO does not utilize gradient ascent. The loss function can be expressed as below:

$$\mathcal{L}_\phi = \mathcal{L}_\phi(\mathcal{D}_r) + \mathcal{L}_\phi(\mathcal{D}_{f,idk}) \tag{6}$$

Unlearning configurations. For all unlearning methods, we conduct optimization of the corresponding loss over 20 epochs. In scenarios where support from the retained set is utilized, an epoch is defined as one complete cycle through the entire forget set, using no more than the same number of samples from the retained set. We employ the AdamW optimizer with a warm-up phase during the first epoch and maintain an effective batch size of 4 for all unlearning algorithms. Regarding the learning rate, we experiment with various settings to identify the optimal rate for each setup and will report these findings individually. We evaluate all baseline methods using the most widely adopted available language models, specifically Llama2-7B [60], and Mistral-7B [29]. All unlearning experiments are run using 1 NVIDIA A40 GPU, and the running time depends on the size and the algorithms.

6 Results

In this section, we demonstrate the benchmark results for each structural unlearning task for each unlearning method described above. We examined learning rates between 5×10^{-6} and 5×10^{-5} during the unlearning process and found that the performances of all unlearning methods are sensitive to the selection of learning rate, albeit to varying degrees. Hence, careful learning rate tuning for each unlearning task is necessary to strike the best balance between unlearning performance and the retained model utility. Since our aim is to demonstrate the impact of dataset topology rather than the unlearning methods *per se*, we presented results with learning rates that generally provide a good balance between unlearning effectiveness and retained model utility. Figure 5 shows the ROUGE1 Score of all unlearning methods under three different unlearning scenarios with both the Llama2-7B and the Mistral-7B model. Table 1 shows the unlearning results with different interconnectivity on Sample Dataset 1. Each experiment is repeated three times, and both average and standard deviation are reported in the exact experiment results. Additional results are included in the Appendix B.

Q1: How will data inter-connectivity impact the unlearning performance?

Answer: As shown by plot(a) and (b) in Figure 5 (and Table 1 and 2), the ROUGE1 Score for the forget set when unlearning the more inter-connected edge AB is higher than when unlearning the less inter-connected edge AC across all unlearning methods. The results indicate that **the greater the degree of inter-connectivity a data point possesses, the more challenging it becomes to unlearn the data.** The difference is particularly pronounced when unlearning with the GA and GD methods, where the ROUGE1 Score for the forget set when unlearning the AB edge is 1.9x and 1.8x higher than unlearning the AC edge respectively. Despite a smaller impact from data inter-connectivity, unlearning is observed to be less effective using the KL method. This phenomenon is likely attributable to the design of KL, which, through reducing distribution shifts before and after the forgetting process, not only encompasses relational information between data points thereby making it more difficult to forget less inter-connected information (such as the *AC* edge), but also makes the forgetting of other data points harder. DPO method demonstrates improved resilience to data inter-connectivity despite being highly sensitive to learning rate changes. One reason for the observed resilience could be that while the other three methods involve gradient ascent on the forget samples, DPO continues to perform gradient descent, albeit following the gradient of the forget set paired with 'I don't know' responses. This suggests that gradient ascent may be less robust for structural unlearning. For all methods, enhancing the forgetting of more inter-connected data points can be achieved by increasing the learning rate. However, this process necessitates meticulous, case-by-case tuning to prevent significant degradation in the retained model's performance, thereby limiting its scalability.

As explained in 4, Sample Dataset 1 has a condensed and symmetric structure which is handy in implementing multi-faceted unlearning but does not yield topological variations when sampling varying unlearning edges. We extend beyond the local view of data inter-connectivity by randomly sampling an edge to forget for each sub-graph in Sample Dataset 2 and measuring the averaged result. As shown in Figure 6, the average ROUGE1 Score for the forget set of the two models, using the GA method, increases from 0.349 to 0.490 as the average data inter-connectivity (i.e. knowledge density) increases from the sparse sub-graph to the dense sub-graph. Similar upward trends have been observed with the GD and the KL methods. The results not only verified the local view of data inter-connectivity revealed using Sample Dataset 1, but also indicate that **the greater density a knowledge graph has, the more challenging it becomes to unlearn the data within the graph.** Similar to the evaluation results with Sample Dataset 1, DPO method appear more robust to knowledge graphs with varying knowledge density, evidenced by relatively stable ROUGE1 Score across sub-graphs in Sample Dataset 2 where the difference is within the statistical margin of error.

Q2: How will the size of the forget set affect overall unlearning performance?

Answer: As we can see in plot (c) and (d) in Figure 5, at the constant learning rate, the ROUGE1 Scores for the forget set demonstrate significant drop for all unlearning methods. This reveals that **the 'batched unlearning' size can significantly affect the unlearning outcomes across all unlearning methods.** Additionally, while an increase in the forget set size effectively erases the targeted data, it may also adversely impacts the fine-tuning process for the retained dataset, demonstrated by the collapse of the ROUGE1 Scores for the retained set for a number of unlearning scenarios in the figure. More critically, it can disastrously render the pre-trained model non-functional or significantly impaired. This detrimental effect is especially pronounced in the given example, where forgetting three edges constitutes a material 15% reduction of the total dataset. To maintain a desirable balance between the forgetting performance and retained model's utility, one can manually test and adjust the learning rate downward for unlearning requests involving larger batch sizes. However, this approach clearly lacks scalability. Future methods will need to enhance flexibility to accommodate varying sizes of forgetting datasets effectively.

Q3: When the forgetting data skewed towards a specific type, will it result in a more pronounced deterioration in the retained model's performance on that same data type as compared to other data types?

Answer: As shown by plot (e) and (f) in Figure 5, ROUGE1 Scores for the independent retained set of sales contracts are lower

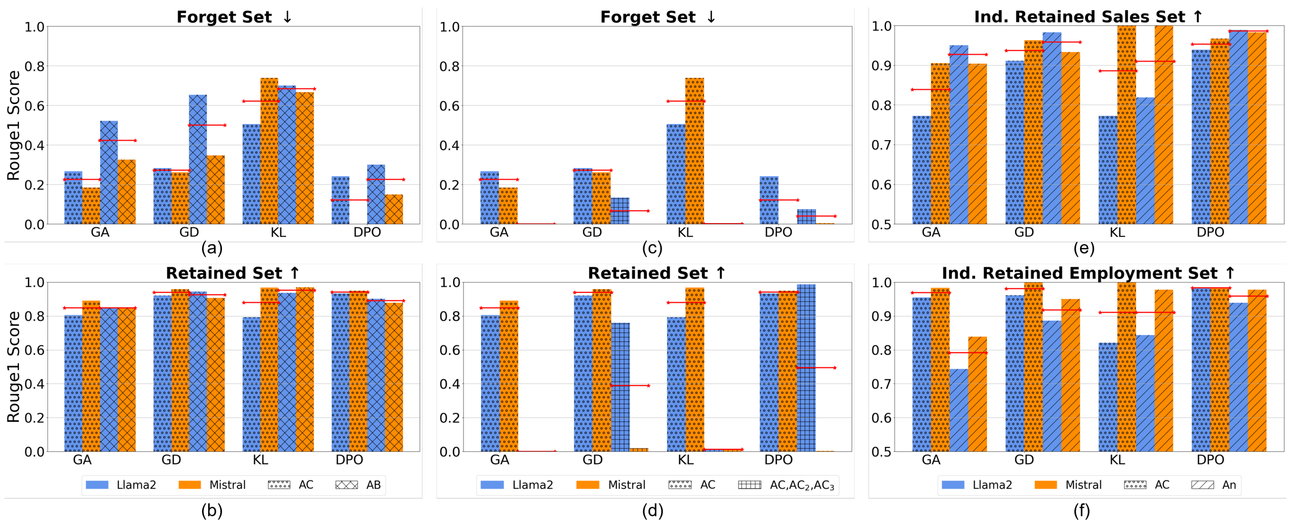

Figure 5: Comparison of ROUGE1 Score on various unlearning scenarios with both Llama2-7B and Mistral-7B model. Plots show the ROUGE1 Score on both forget set (↓ the lower the better) and retained set (↑ the higher the better). (a,b) demonstrates the difference between removing a highly inter-connected branch (the *AB* edge) and a lesser connected leaf (the *AC* edge). (c,d) shows the comparison between removing a single sample (*AC* edge) and multiple samples (*AC*, AC_2, AC_3). (e,f) shows the comparison between removing a sales contract (*AC*) and an employment contract (*An*) for the independent retained set of each data type. The red line indicates the average of each setup of the two models.

Table 1: Unlearning results with Different Interconnectivity on Sample Dataset 1.

Llama2-7B		Rouge Score			MRR			Top Hit Ratio		
Forget Edge	Forget Method	Forget Set	Retained Set	Factual Set	Forget Set	Retained Set	Factual Set	Forget Set	Retained Set	Factual Set
AB	GA	0.521 ± 0.050	0.845 ± 0.043	0.871 ± 0.008	0.288 ± 0.004	0.310 ± 0.004	0.224 ± 0.009	0.766 ± 0.024	0.759 ± 0.004	0.580 ± 0.005
	GD	0.654 ± 0.029	0.944 ± 0.003	0.866 ± 0.003	0.277 ± 0.008	0.292 ± 0.002	0.233 ± 0.011	0.731 ± 0.021	0.733 ± 0.006	0.577 ± 0.010
	KL	0.700 ± 0.050	0.936 ± 0.036	0.960 ± 0.004	0.355 ± 0.014	0.370 ± 0.017	0.205 ± 0.009	0.839 ± 0.016	0.831 ± 0.010	0.569 ± 0.008
	DPO	0.300 ± 0.000	0.902 ± 0.023	0.891 ± 0.006	0.261 ± 0.011	0.295 ± 0.002	0.221 ± 0.001	0.600 ± 0.003	0.690 ± 0.009	0.589 ± 0.002
AC	GA	0.267 ± 0.029	0.805 ± 0.005	0.846 ± 0.010	0.179 ± 0.001	0.271 ± 0.004	0.251 ± 0.011	0.606 ± 0.004	0.735 ± 0.006	0.579 ± 0.008
	GD	0.283 ± 0.029	0.920 ± 0.005	0.850 ± 0.004	0.168 ± 0.002	0.261 ± 0.002	0.250 ± 0.005	0.590 ± 0.009	0.728 ± 0.008	0.581 ± 0.003
	KL	0.505 ± 0.256	0.793 ± 0.165	0.955 ± 0.005	0.357 ± 0.031	0.407 ± 0.044	0.206 ± 0.002	0.694 ± 0.073	0.784 ± 0.047	0.568 ± 0.002
	DPO	0.242 ± 0.000	0.933 ± 0.008	0.881 ± 0.001	0.159 ± 0.001	0.277 ± 0.001	0.221 ± 0.001	0.425 ± 0.003	0.658 ± 0.009	0.576 ± 0.001
Mistral-7B										
AB	GA	0.325 ± 0.029	0.851 ± 0.010	0.917 ± 0.005	0.166 ± 0.014	0.269 ± 0.004	0.566 ± 0.004	0.500 ± 0.013	0.678 ± 0.007	0.790 ± 0.002
	GD	0.347 ± 0.075	0.906 ± 0.021	0.918 ± 0.002	0.163 ± 0.002	0.285 ± 0.004	0.551 ± 0.005	0.524 ± 0.044	0.713 ± 0.013	0.781 ± 0.002
	KL	0.667 ± 0.052	0.969 ± 0.014	0.961 ± 0.003	0.293 ± 0.028	0.316 ± 0.008	0.726 ± 0.005	0.737 ± 0.009	0.784 ± 0.004	0.892 ± 0.003
	DPO	0.150 ± 0.050	0.878 ± 0.019	0.945 ± 0.005	0.081 ± 0.019	0.271 ± 0.004	0.624 ± 0.013	0.187 ± 0.043	0.631 ± 0.019	0.834 ± 0.004
AC	GA	0.184 ± 0.058	0.890 ± 0.010	0.916 ± 0.004	0.134 ± 0.054	0.236 ± 0.004	0.513 ± 0.006	0.252 ± 0.079	0.597 ± 0.025	0.755 ± 0.008
	GD	0.261 ± 0.077	0.959 ± 0.010	0.929 ± 0.004	0.149 ± 0.006	0.266 ± 0.005	0.505 ± 0.004	0.293 ± 0.047	0.635 ± 0.016	0.748 ± 0.003
	KL	0.739 ± 0.010	0.968 ± 0.008	0.966 ± 0.003	0.246 ± 0.002	0.279 ± 0.008	0.700 ± 0.012	0.727 ± 0.019	0.757 ± 0.008	0.877 ± 0.004
	DPO	0.000 ± 0.000	0.949 ± 0.004	0.949 ± 0.003	0.018 ± 0.001	0.269 ± 0.004	0.598 ± 0.001	0.125 ± 0.000	0.676 ± 0.001	0.820 ± 0.001

than those of employment contracts when the unlearning edge, *AC*, is a sales contract. The opposite behavior is observed when the unlearning edge is switched to *An*, an employment contract – with which ROUGE1 Scores for the independent retained set of employment contracts are lower than those of sales contracts. The results indicate that **unlearning a specific type of data may result in the model's performance deteriorating more on the same type of data than on different types of data.** Figure 5 also shows that the difference in unlearning outcomes is more pronounced when employing the GA method compared to other methods, indicating a greater lack of robustness in the GA method when handling this kind of unlearning task.

Q4: Will the choice of pre-trained model affect the unlearning performance?

Answer: When evaluating the impact of new unlearning tasks on pre-trained models of comparable size, such as Llama2-7B and Mistral-7B, we found that the outcomes of unlearning can vary depending on the specific pre-trained model chosen.

We compare the unlearning outcome across all scenarios for two pre-trained models – Llama2-7B (blue bars) and Mistral-7B (orange

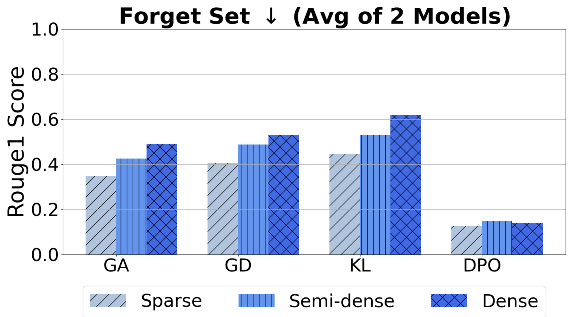

Figure 6: Forget Set Performance of Varying Knowledge Density (avg of the two models).

bars) – in Figure 5. As illustrated by plot(a) – (d) in Figure 5, topological features such as data inter-connectivity and the size of the forget set generally have a similar impact on both the Mistral-7B and Llama-7B models. However, with roughly the same ROUGE1 Score for the retained set, Mistral-7B model generally demonstrates higher unlearning effectiveness with the GA, GD and DPO methods, evidenced by lower ROUGE1 Scores for the forget set. However, reduced unlearning effectiveness using the KL method with the Mistral-7B model has been observed. This is likely attributable to the Mistral-7B model's stronger inherent resistance to distribution shifts before and after the unlearning process, a characteristic designed into the KL method. Additionally, due to the Mistral-7B model's stronger resistance to distribution shifts, the impact of data inter-connectivity when using the KL methods is not significant (the disparity shown in Figure 5 falls within the statistical margin of error). Furthermore, while the unlearning outcomes for data skewed towards a specific type show similar behavioral tendencies in both the Mistral-7B and Llama2-7B models, the differences are less pronounced in the Mistral-7B model. This is likely due to the higher unlearning effectiveness of the Mistral-7B model, where achieving comparable unlearning outcomes appears to degrade the performance on the retained set to a lesser extent. Consequently, this reduces the impact of data type on the performance of the retained dataset in the Mistral-7B model.

Our findings indicate that the choice of pre-trained model will affect unlearning performance, with the degree of impact varying based on several factors, including the unlearning method, learning rate, and dataset used, among others. Consequently, we advocate for the inclusion of unlearning performance evaluation, particularly their robustness in relation to real-world data topology, within the broader assessment frameworks for pre-trained LLMs.

7 Discussion and Roadmap

LLMs unlearning is an evolving area that requires significant enhancements across various facets. Our new dataset compilation pipeline, sample datasets, and benchmark results highlight several challenges and future research directions in this domain.

In summary, *PISTOL* offers an easy-to-use and versatile pipeline for synthesizing knowledge-graph based datasets that uniquely captures the structural relationships among entities for the evaluation of structural LLM unlearning. Our sample datasets, compiled using *PISTOL*, not only show that **current unlearning methods are highly sensitive to hyperparameters, leading to unstable outcomes**, but also demonstrate **a lack robustness when applied to structured data**. Our findings reveal that current unlearning methods struggle to handle scenarios characterized by (1) varying levels of inter-connectivity among data points, (2) larger sizes of data sets to be forgotten, and (3) unlearning tasks targeting a specific type of data. The challenges encountered in our experiments underscore the critical need for future research aimed at developing more effective and robust machine unlearning methods for LLMs.

With respect to future research direction, limited existing research on using model editing for unlearning, as discussed in Section 2, suggests that this could be a promising approach. However, this approach also necessitates a more precise and widely accepted definition of what constitutes sufficient data deletion, as opposed to mere data editing. Additionally, integrating unlearning with retrieval augmented generation (RAG) techniques could enhance model updates, particularly in fields where accuracy and timeliness are crucial, such as news, medicine, or law. Furthermore, in contrast to the traditional LLM training paradigm,

Moreover, in contrast to the traditional LLM training paradigm and inspired by federate learning techniques [6, 33, 52, 73], federated LLM pre-training and fine-tuning [26, 55, 72] are gaining increasing attention as this approach not only personalizes the model but also leverages private data to enhance the capabilities of pre-trained LLMs. Importantly, within FL scenarios, the concept of unlearning becomes crucial as each participating client retains the right to forget, necessitating the development of federated unlearning methods [22, 38]. These methods must address the unique challenge of removing a client's influence on the trained model—an aspect of critical importance in federated environments. Presently, federated unlearning techniques primarily address standard classification tasks without considering structural relationships between client datasets. Extending these unlearning methods to efficiently, effectively, and robustly handle structured datasets in decentralized training environments remains a vital area for future research.

Furthermore, Differential Privacy (DP) was initially designed to protect individual data privacy and has since been adapted to safeguard privacy at the client level [13, 51, 65, 69]. Training models under DP often involve a trade-off between the privacy budget and model utility . Interestingly, DP also imparts a degree of generalizability to the model by diluting the influence of individual data points or clients. Exploring how DP can be integrated with unlearning methods to enhance the model's forgetting capabilities without compromising utility presents an intriguing research opportunity. This integration could potentially lead to more robust FL frameworks that uphold both privacy and performance, aligning with the evolving needs of secure and responsible AI development.

Lastly, our results also reveal that the choice of pre-trained model does influence unlearning performance. We therefore recommend, for the first time, the incorporation of unlearning performance evaluation, especially its robustness in relation to real-world data topology, into the broader assessment frameworks for pre-trained LLMs.

References

[1] [n. d.]. Right to erasure ('right to be forgotten'). The New York Times,https://gdpr-info.eu/art-17-gdpr/.

[2] Sara Abdali, Richard Anarfi, CJ Barberan, and Jia He. 2024. Decoding the AI Pen: Techniques and Challenges in Detecting AI-Generated Text. *arXiv preprint arXiv:2403.05750* (2024).

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. 2023. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security* 6, 1 (2023), 1–52.

[5] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[6] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, et al. 2020. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390* (2020).

[7] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 141–159.

[8] 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. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[9] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45.

[10] Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150* (2023).

[11] Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Adelani, Pontus Lars Erik Saito Stenetorp, Sebastian Riedel, and Mikel Artetxe. 2023. Improving language plasticity via pretraining with active forgetting. *Advances in Neural Information Processing Systems* 36 (2023), 31543–31557.

[12] Yihong Chen, Pasquale Minervini, Sebastian Riedel, and Pontus Stenetorp. 2021. Relation Prediction as an Auxiliary Training Objective for Improving Multi-Relational Graph Representations. In *3rd Conference on Automated Knowledge Base Construction*. https://openreview.net/forum?id=Qa3uS3H7-Le

[13] Giovanni Cherubin, Boris Köpf, Andrew Paverd, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2024. Closed-Form Bounds for DP-SGD against Record-level Inference. *arXiv preprint arXiv:2402.14397* (2024).

[14] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[15] Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing Factual Knowledge in Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 6491–6506.

[16] Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680* (2022).

[17] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 5484–5495.

[18] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9304–9312.

[19] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. 2019. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030* (2019).

[20] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736* (2023).

[21] Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024. Model Editing at Scale leads to Gradual and Catastrophic Forgetting. *arXiv preprint arXiv:2401.07453* (2024).

[22] Anisa Halimi, Swanand Kadhe, Ambrish Rawat, and Nathalie Baracaldo. 2022. Federated unlearning: How to efficiently erase a client in fl? *arXiv preprint arXiv:2207.05521* (2022).

[23] Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2021. Do language models have beliefs? methods for detecting, updating, and visualizing model beliefs. *arXiv preprint arXiv:2111.13654* (2021).

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[25] Taojun Hu and Xiao-Hua Zhou. 2024. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions. *arXiv preprint arXiv:2404.09135* (2024).

[26] Alex Iacob, Lorenzo Sani, Bill Marino, Preslav Aleksandrov, and Nicholas Donald Lane. 2024. Worldwide Federated Training of Language Models. *arXiv preprint arXiv:2405.14446* (2024).

[27] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

[28] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2022. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504* (2022).

[29] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[30] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*. 12–24.

[31] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2024. Towards Unbounded Machine Unlearning. *Advances in Neural Information Processing Systems* 36 (2024).

[32] Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. 2018. Canonical tensor decomposition for knowledge base completion. In *International Conference on Machine Learning*. PMLR, 2863–2872.

[33] Royson Lee, Minyoung Kim, Da Li, Xinchi Qiu, Timothy Hospedales, Ferenc Huszár, and Nicholas Lane. 2024. FedL2P: Federated Learning to Personalize. *Advances in Neural Information Processing Systems* 36 (2024).

[34] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[35] Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 150–157. https://aclanthology.org/N03-1020

[36] Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. Barcelona, Spain, 605–612. https://doi.org/10.3115/1218955.1219032

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

[38] Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. 2020. Federated unlearning. *arXiv preprint arXiv:2012.13891* (2020).

[39] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R Varshney, et al. 2024. Rethinking Machine Unlearning for Large Language Models. *arXiv preprint arXiv:2402.08787* (2024).

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

[41] Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. 2023. Mini-Model Adaptation: Efficiently Extending Pretrained Models to New Languages via Aligned Shallow Training. In *Findings of the Association for Computational Linguistics: ACL 2023*. 5474–5490.

[42] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems* 35 (2022), 17359–17372.

[43] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2022. Mass-Editing Memory in a Transformer. In *The Eleventh International Conference on Learning Representations*.

[44] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast Model Editing at Scale. In *International Conference on Learning Representations*.

[45] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *International Conference on Machine Learning*. PMLR, 15817–15831.

[46] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: Measuring ChatGPT political bias. *Public Choice* (2023), 1–21.

[47] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035* (2023).

[48] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine

unlearning. *arXiv preprint arXiv:2209.02299* (2022).

[49] Vaidehi Patil, Peter Hase, and Mohit Bansal. 2023. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *arXiv preprint arXiv:2309.17410* (2023).

[50] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*. Association for Computational Linguistics, Online, 46–54. https://www.aclweb.org/anthology/2020.emnlp-demos.7

[51] Xinchi Qiu, Ilias Leontiadis, Luca Melis, Alex Sablayrolles, and Pierre Stock. 2023. EXACT: Extensive Attack for Split Learning. *arXiv preprint arXiv:2305.12997* (2023).

[52] Xinchi Qiu, Heng Pan, Wanru Zhao, Chenyang Ma, Pedro PB Gusmao, and Nicholas D Lane. 2023. vfedsec: Efficient secure aggregation for vertical federated learning via secure layer. *arXiv preprint arXiv:2305.16794* (2023).

[53] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2024).

[54] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics* 8 (2021), 842–866.

[55] Lorenzo Sani, Alex Iacob, Zeyu Cao, Bill Marino, Yan Gao, Tomas Paulik, Wanru Zhao, William F Shen, Preslav Aleksandrov, Xinchi Qiu, et al. 2024. The Future of Large Language Model Pre-training is Federated. *arXiv preprint arXiv:2405.10853* (2024).

[56] Adam Schiff. [n. d.]. A BILL To require a notice be submitted to the Register of Copyrights with respect to copyrighted works used in building generative AI systems, and for other purposes. https://schiff.house.gov/imo/media/doc/the_generative_ai_copyright_disclosure_act.pdf

[57] Johannes Schneider and Michalis Vlachos. 2021. Explaining neural networks by decoding layer activations. In *Advances in Intelligent Data Analysis XIX: 19th International Symposium on Intelligent Data Analysis, IDA 2021, Porto, Portugal, April 26–28, 2021, Proceedings 19*. Springer, 63–75.

[58] Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov, and Artem Babenko. 2019. Editable Neural Networks. In *International Conference on Learning Representations*.

[59] Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. 2022. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX Security Symposium (USENIX Security 22)*. 4007–4022.

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

[61] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 2071–2080. https://proceedings.mlr.press/v48/trouillon16.html

[62] 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* (2023).

[63] Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. *arXiv preprint arXiv:2311.17391* (2023).

[64] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 4003–4012. https://aclanthology.org/2020.lrec-1.494

[65] Lukas Wutschitz, Huseyin A Inan, and Andre Manoel. 2022. dp-transformers: Training transformer models with differential privacy.

[66] Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do Large Language Models Latently Perform Multi-Hop Reasoning? *arXiv preprint arXiv:2402.16837* (2024).

[67] Yunzhi Yao, Shaohan Huang, Li Dong, Furu Wei, Huajun Chen, and Ningyu Zhang. 2022. Kformer: Knowledge injection in transformer feed-forward layers. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 131–143.

[68] Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *arXiv preprint arXiv:2310.10683* (2023).

[69] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500* (2021).

[70] Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. 2023. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *arXiv preprint arXiv:2307.03941* (2023).

[71] Jinghan Zhang, Junteng Liu, Junxian He, et al. 2023. Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems* 36 (2023), 12589–12610.

[72] Wanru Zhao, Royson Lee, Yihong Chen, Xinchi Qiu, Yan Gao, Hongxiang Fan, and Nicholas Donald Lane. 2023. Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages. In *The Twelfth International Conference on Learning Representations*.

[73] Wanru Zhao, Xinchi Qiu, Javier Fernandez-Marques, Pedro PB de Gusmão, and Nicholas D Lane. 2022. Protea: Client profiling within federated systems using flower. In *Proceedings of the 1st ACM Workshop on Data Privacy and Federated Learning Technologies for Mobile Edge Network*. 1–6.

Appendix

A Sample Dataset

Sample of Sales of Goods Contract

SALES OF GOODS CONTRACT

1. PARTIES

- This Sales Contract (hereinafter referred to as the "**Contract**") is entered into on [•] (the "**Effective Date**") by and between [•] with an address of [•] (the "**Seller**") and [•] with an address of [•] (the "**Customer**") (collectively referred to as the "**Parties**").

2. GOODS AND PRICE

- The goods that the Seller is selling to the Customer are enlisted below with their quantities (hereinafter referred to as the "**Goods**").
  - Goods: [•]
  - Quantity: [•]
  - Price per unit: [•]
  - Total price: [•]

3. PAYMENTS

- The Seller shall provide the Customer with an invoice no later than [•] days after the time of the delivery.
- All invoices are to be paid in full within [•] days. Any balances not paid within [•] days will be subject to a [•]% late payment penalty.

4. DELIVERY AND SHIPPING

- The delivery of the goods (the "**Delivery**") will be at the location [•].
- The shipping method will be decided by the [•]. [•] will be responsible for the costs of the shipment.

5. WARRANTIES

- **General Warranty:** The Seller hereby warrants to the Customer that the Goods shall be free from defects in materials and workmanship under normal use and service for a period of [•] years from the date of delivery (the "**Warranty Period**"). The Seller affirms that it has good title to the Goods free and clear of any liens and encumbrances and has the right to sell the Goods to the Customer.
- **Remedy for Breach of Warranty:** In the event of a breach of this warranty, the Customer must notify the Seller in writing within of [•] days of discovering the defect. Upon receiving such notification, the Seller shall, at its sole option, (i) repair or replace the defective Goods at no additional charge to the Customer, or (ii) refund the purchase price paid for the defective Goods, provided that the Goods are returned to the Seller, if so requested. The choice of remedy shall be at the Customer's discretion if repair or replacement does not remedy the defect within a reasonable time.
- **Exclusions from Warranty:** This warranty does not apply to any damage or defect resulting from misuse, abuse, neglect, alterations, unauthorized repairs, modifications, accidents, or natural wear and tear. The Seller's obligation under this warranty is limited to the repair, replacement, or refund as provided under this section and does not cover any other costs such as the cost of removal and reinstallation of Goods, loss of use, loss of profit, or other incidental or consequential damages.
- **No Other Warranties:** Except for the warranty set forth herein, the Seller disclaims all other warranties, express or implied, including, but not limited to, any implied warranties of merchantability or fitness for a particular purpose. The Seller's liability under this warranty shall be limited to the repair, replacement, or refund as specified herein, and in no event shall exceed the purchase price of the defective Goods.

- **Survival:** This warranty shall survive the delivery, inspection, acceptance, and payment of and for the Goods and shall inure to the benefit of the Customer and its successors and assigns.

6. INSPECTION

- Hereby, the Customer acknowledges that it has relied solely on the investigations, examinations, and inspections that the Customer has chosen to make and that the Seller has afforded the Customer the opportunity for full and complete investigations, examinations, and inspections.

7. RISK OF LOSS AND TITLE

- The risk of loss or damage for the goods will be on the Seller until the goods pass upon delivery to the Customer or its designee. The Title of the goods will also remain with the Seller until the goods pass upon delivery to the Customer or its designee.

8. DELAY OR FAILURE TO PERFORM AND FORCE MAJEURE

- Under no circumstances will the Seller be held liable to the Customer for any delay that may occur, non-delivery or an arising fault of this Agreement that may be due to any labour dispute, shortage in transportation, delay or shortage of materials to produce the Goods, fires, accidents, Acts of God, or any other causes outside Seller's control. The Seller will notify the Customer immediately upon realization that it will not be able to deliver the Goods as promised. Upon such notice, either Party may terminate this Agreement.

9. COOLING-OFF PERIOD

- Either Party may terminate this Agreement, for any reason, within [•] days following the Effective Date of this Agreement ('**Cooling-Off Period**'). Termination during this Cooling-Off Period must be communicated in writing to the other Party. Following the expiration of the Cooling-Off Period, no Party shall have the right to terminate this Agreement on the basis of the Cooling-Off Period provisions.

10. LIMITATION OF LIABILITY

- Under no circumstances will the Seller be liable for any indirect, special, consequential, or punitive damages (including lost profits) arising out of or relating to this Agreement or the transactions it contemplates (whether for breach of contract, tort, negligence, or other form of action).

11. SEVERABILITY

- In the event that any provision of this Agreement is found to be void and unenforceable by a court of competent jurisdiction, then the remaining provisions will remain in force in accordance with the Parties' intention.

12. ENTIRE AGREEMENT

- This Agreement contains the entire agreement and understanding among the Parties hereto with respect to the subject matter hereof, and supersedes all prior agreements, understandings, inducements and conditions, express or implied, oral or written, of any nature whatsoever with respect to the subject matter hereof. The express terms hereof control and supersede any course of performance and/or usage of the trade inconsistent with any of the terms hereof.

13. GOVERNING LAW

- This Agreement shall be governed by and construed in accordance with the laws of [•].

The Parties hereby agree to the terms and conditions set forth in this Agreement and such is demonstrated throughout their signatures below.

Sample QAs of the Sales of Goods Contract

- Q1: What was the effective date of the contract between [*seller name*] and [*customer name*]?
- Q2: What was the name of the seller in the contract with [*customer name*] as of [*effective date*]?
- Q3: What was the address of [*seller name*] in the contract with [*customer name*]?
- Q4: What was the name of the customer in the contract with [*seller name*] as of [*effective date*]?
- Q5: What was the address of [*customer name*] in the contract with [*seller name*]?
- Q6: What was the good that the seller was selling to the customer based on the contract between [*seller name*] and [*customer name*]?
- Q7:What was the quantity of the good being sold based on the contract between [*seller name*] and [*customer name*]?
- Q8: What was the unit price in dollars of the good being sold based on the contract between [*seller name*] and [*customer name*]?
- Q9: What was the total price in dollars of the good being sold based on the contract between [*seller name*] and [*customer name*]?
- Q10: By how many days after the delivery time must the seller provide the customer with an invoice based on the contract between [*seller name*] and [*customer name*]?
- Q11: Within how many days must the invoice be paid in full based on the contract between [*seller name*] and [*customer name*]?
- Q12: After how many days would unpaid balances incur a late payment penalty based on the contract between [*seller name*] and [*customer name*]?
- Q13: What was the late payment interest rate based on the contract between [*seller name*] and [*customer name*]?
- Q14: What was the address of delivery based on the contract between [*seller name*] and [*customer name*]?
- Q15: Who would decide the shipping method based on the contract between [*seller name*] and [*customer name*]?
- Q16: Who would be responsible for the costs of the shipment based on the contract between [*seller name*] and [*customer name*]?
- Q17: What was the duration of the general warranty period in years based on the contract between [*seller name*] and [*customer name*]?
- Q18: Within how many days of discovering a defect must the customer notify the seller in writing in the event of a breach of warranty based on the contract between [*seller name*] and [*customer name*]?
- Q19: What was the duration of the cooling-off period in days based on the contract between [*seller name*] and [*customer name*]?
- Q20: Which jurisdiction's laws govern the contract between [*seller name*] and [*customer name*]?

Sample of Employment Contract

EMPLOYMENT CONTRACT

1. PARTIES

- (the "**Employer**") with its principal place of business located at [•] ("**Employer's Business Address"**) agrees to employ [•] (the "**Employee**") who lives at [•] ("**Employee's Residential Address**") and the Employee agrees to be employed, on the terms and conditions set out in this Contract, and in the accompanying Addendum (together, the "**Agreement**").

2. START AND LENGTH OF EMPLOYMENT

- The Employee will start employment on [•] ("**Start Date**").
- The Employer shall employ the Employee for [•] months ("**Length of Employment**"), however, the Employer and the Employee may change the Length of Employment in accordance with Clause 12 of this Contract.

3. JOB TITLE AND DUTIES

- The Employee shall be employed as [•] ("**Position**"). The Employee shall perform the duties as described in the accompanying Addendum, and any other duties reasonably assigned by the Employer.

4. PLACE OF WORK

- The Employee shall work at [•] ("**Address of Work Location**"). The Employee shall not be required to work at a different location unless the Employee consents in writing to such an arrangement. Any such employment shall be on the same terms and conditions as this Agreement.

5. WORKING HOURS

- The Employee's normal days of work are Monday to Friday ("**Normal Work Days**") and the Employee's normal hours of work are [•] to [•] ("**Normal Work Hours**") (together, the "**Work Week**").

6. PAY

- The Employer shall pay the Employee $[•] ("**Rate of Basic Pay**") per hour.
- The Employer shall pay the Employee in [•] instalments.

7. BENEFITS

- The Employee shall be entitled to participate in [•] offered by the Employer, subject to the terms and conditions of those plans.

8. HOLIDAYS

- The Employer shall provide the Employee with [•] days of paid holiday leave ("**Holiday Leave**") per year, plus the public holidays.
- The Employee shall provide the Employer two weeks' notice of any Holiday Leave, and the Employer may only refuse Holiday Leave in exceptional circumstances. The Employer shall pay the Employee for any unused Holiday Leave at the earlier of: (i) the end of each year, or (ii) the end of the Employee's employment.

9. CONFIDENTIALITY

- The Employee agrees that during the term of employment and for the first [•] months thereafter, he/she will not disclose any confidential information pertaining to the business of the Employer to any person not authorized by the Employer to receive such information.

10. WORK CONDITIONS

- The Employer shall ensure the Employee is appropriately instructed and trained in relation to tasks that the Employee will carry out. The Employer shall provide a safe and healthy work environment and shall not require the Employee to do work that subjects the Employee to health or safety hazards.

11. SICK PAY AND ABSENCE

- The Employee shall notify the Employer if he or she is going to be absent from work because of sickness or injury. The Employer shall not require the Employee to work when sick or injured.
- In each year of employment, the Employee shall be entitled to receive the Basic Rate of Pay (as if he/she had worked the Normal Work Hours) per day for the first [•] days of absence from work due to sickness or injury ("**Paid Sick Leave**").

12. TERMINATION

- The Employee and Employer shall each provide the other with [•] weeks' written notice of termination in a language the Employee understands.
- The Employee and Employer may agree that the Employer pay the Employee for this notice period instead of requiring the Employee to work. In exceptional circumstances, as defined in the Addendum, notice of termination is not required.
- On termination, the Employee shall return to the Employer all Employer property, and the Employer shall pay immediately all monies due under this Agreement to the Employee.

13. NON-COMPETE

- During the term of employment and for [•] months after the termination of employment, the Employee agrees not to engage in any business activities or employment with a competitor or in any capacity that directly competes with the Employer's business within the United States.
- This restriction applies to similar products, services, or industry sectors in which the Employer operates. The Employee acknowledges that such competition could harm the Employer's business interests and agrees to refrain from such activities to protect the Employer's legitimate business interests.

14. CHANGES TO EMPLOYMENT TERMS

- This Contract and the attached Addendum make up the entire Agreement relating to the Employee's employment. The Employer shall not make any changes to this Agreement without the Employee's written consent. The Employer shall provide [•] weeks' written notice of any proposed changes in a language the Employee understands, and the Employer shall permit the Employee to ask questions about such changes.

15. ENTIRE AGREEMENT

- This Agreement and the attached Addendum contain the entire agreement between the parties. The Employee acknowledges that he/she has not relied on any oral or written representations made by the Employer or its employees or agents.

16. GOVERNING LAW

- This Agreement and any dispute or claim arising out of or in connection with it or its subject matter or formation (including non-contractual disputes or claims) shall be governed by and construed in accordance with the laws of [•] ("**Governing Law**").

I acknowledge that I have read this Contract and the Addendum to this Contract; I understand and accept the terms and conditions set out within it, and that this Contract, together with the Addendum, form the Agreement of Employment.

Sample QAs of the Employment Contract

- Q1: What was the name of the employer in the employment contract with [*employee name*], which started from [*start date*]?
- Q2: What was the principal business location of [*employer name*] based on the contract between [*employer name*] and [*employee name*]?
- Q3: What was the name of the employee in the employment contract with [*employer name*], which started from [*start date*]?
- Q4: What was the address of [*employee name*] based on the contract between [*employer name*] and [*employee name*]?
- Q5: What was the start date based on the contract between [*employer name*] and [*employee name*]?
- Q6: For how many months will the employer employ the employee based on the contract between [*employer name*] and [*employee name*]?
- Q7: What was the job position based on the contract between [*employer name*] and [*employee name*]?
- Q8: What was the work location based on the contract between [*employer name*] and [*employee name*]?
- Q9: At what hour did the workday start based on the contract between [*employer name*] and [*employee name*]?
- Q10: At what hour did the workday finish based on the contract between [*employer name*] and [*employee name*]?
- Q11: What was the hourly basic pay in dollars based on the contract between [*employer name*] and [*employee name*]?
- Q12: What was the frequency of salary payment based on the contract between [*employer name*] and [*employee name*]?
- Q13: What benefit was provided to the employee based on the contract between [*employer name*] and [*employee name*]?
- Q14: How many days of paid holiday leave were provided to the employee based on the contract between [*employer name*] and [*employee name*]?
- Q15: For how many months after the employment ends was the employee prohibited from disclosing any confidential information based on the contract between [*employer name*] and [*employee name*]?
- Q16: What was the number of days the employee was entitled to Paid Sick Leave in each year of employment based on the contract between [*employer name*] and [*employee name*]?
- Q17: How many weeks' written notice of termination must the employee and employer each provide to the other based on the contract between [*employer name*] and [*employee name*]?
- Q18: For how many months did the non-compete clause cover based on the contract between [*employer name*] and [*employee name*]?
- Q19: How many weeks' written notice must the employer provide before any proposed changes to the terms of employment based on the contract between [*employer name*] and [*employee name*]?
- Q20: Which jurisdiction's laws govern the contract between [*employer name*] and [*employee name*]?

B Additional Experiment Results

In this section, we show in detail all the results of our benchmark. As explained in Section 6, unlearning performance is very sensitive to the choice of learning rate. However, since our aim is to demonstrate the impact of dataset topology rather than the unlearning methods *per se*, we presented results with learning rates that generally provide a good balance between unlearning effectiveness and retained model utility. With Sample Dataset 1 and Llama2-7B model, we chose the learning rate of 2×10^{-5} for the GA, GD, and KL methods and 1.5×10^{-5} for the DPO method. With Sample Dataset 1 and Mistral-7B model, we chose the learning rate of 1×10^{-5} for the GA, GD and DPO methods and 2×10^{-5} for the KL method. With Sample Dataset 2, we chose the learning rate of 1×10^{-5} for the GA, GD and DPO methods and of 2×10^{-5} for the KL method for both models.

The full tables for Figure 5 can be found in Table 1, 3 and 4. The full table for Figure 6 can be found in Table 2.

All the experiments are repeated three times, and the average and standard deviation are reported in the tables.

Table 2: Unlearning results with Different Knowledge Density on Sample Dataset 2

Llama2-7B		Rouge Score		MRR		Top Hit Ratio	
Knowledge Density	Forget Method	Forget Set	Retained Set	Forget Set	Retained Set	Forget Set	Retained Set
Sparse	GA	0.430 ± 0.104	0.994 ± 0.004	0.179 ± 0.039	0.278 ± 0.006	0.516 ± 0.056	0.610 ± 0.009
	GD	0.538 ± 0.144	0.997 ± 0.003	0.202 ± 0.054	0.279 ± 0.003	0.522 ± 0.052	0.608 ± 0.009
	KL	0.197 ± 0.085	0.478 ± 0.092	0.226 ± 0.045	0.366 ± 0.022	0.427 ± 0.137	0.564 ± 0.061
	DPO	0.220 ± 0.091	0.980 ± 0.023	0.158 ± 0.040	0.280 ± 0.005	0.400 ± 0.035	0.599 ± 0.019
Semi-Dense	GA	0.469 ± 0.055	0.999 ± 0.002	0.197 ± 0.039	0.277 ± 0.001	0.544 ± 0.058	0.609 ± 0.006
	GD	0.598 ± 0.074	1.000 ± 0.000	0.240 ± 0.053	0.277 ± 0.001	0.584 ± 0.086	0.609 ± 0.005
	KL	0.258 ± 0.086	0.405 ± 0.141	0.255 ± 0.073	0.333 ± 0.029	0.422 ± 0.114	0.518 ± 0.086
	DPO	0.220 ± 0.055	0.996 ± 0.004	0.205 ± 0.042	0.279 ± 0.001	0.460 ± 0.086	0.608 ± 0.010
Dense	GA	0.576 ± 0.093	0.997 ± 0.003	0.215 ± 0.037	0.276 ± 0.001	0.573 ± 0.062	0.608 ± 0.008
	GD	0.623 ± 0.115	0.999 ± 0.002	0.223 ± 0.032	0.277 ± 0.001	0.583 ± 0.049	0.610 ± 0.008
	KL	0.302 ± 0.080	0.528 ± 0.110	0.268 ± 0.048	0.376 ± 0.005	0.521 ± 0.149	0.593 ± 0.063
	DPO	0.230 ± 0.057	0.998 ± 0.004	0.219 ± 0.053	0.278 ± 0.002	0.470 ± 0.100	0.602 ± 0.006
Mistral-7B							
Sparse	GA	0.268 ± 0.098	0.938 ± 0.027	0.223 ± 0.049	0.311 ± 0.006	0.517 ± 0.100	0.657 ± 0.051
	GD	0.272 ± 0.102	0.957 ± 0.017	0.218 ± 0.048	0.313 ± 0.008	0.520 ± 0.110	0.662 ± 0.044
	KL	0.694 ± 0.177	0.997 ± 0.004	0.290 ± 0.044	0.314 ± 0.009	0.665 ± 0.152	0.735 ± 0.045
	DPO	0.030 ± 0.027	0.976 ± 0.010	0.028 ± 0.013	0.309 ± 0.002	0.142 ± 0.025	0.670 ± 0.017
Semi-dense	GA	0.380 ± 0.087	0.977 ± 0.012	2.201 ± 0.040	0.313 ± 0.012	0.540 ± 0.094	0.669 ± 0.069
	GD	0.377 ± 0.089	0.983 ± 0.007	0.206 ± 0.030	0.313 ± 0.013	0.550 ± 0.099	0.668 ± 0.080
	KL	0.803 ± 0.110	0.999 ± 0.002	0.303 ± 0.041	0.320 ± 0.002	0.721 ± 0.045	0.751 ± 0.004
	DPO	0.075 ± 0.056	0.991 ± 0.004	0.056 ± 0.027	0.306 ± 0.002	0.187 ± 0.068	0.671 ± 0.013
Dense	GA	0.403 ± 0.096	0.961 ± 0.035	0.264 ± 0.079	0.318 ± 0.015	0.566 ± 0.117	0.685 ± 0.031
	GD	0.435 ± 0.079	0.968 ± 0.034	0.273 ± 0.074	0.320 ± 0.013	0.588 ± 0.142	0.683 ± 0.033
	KL	0.935 ± 0.038	1.000 ± 0.000	0.318 ± 0.033	0.319 ± 0.003	0.764 ± 0.038	0.757 ± 0.006
	DPO	0.050 ± 0.061	0.993 ± 0.006	0.039 ± 0.028	0.307 ± 0.005	0.158 ± 0.095	0.667 ± 0.017

Table 3: Unlearning results with different quantity

Llama2-7B		Rouge Score			MRR			Top Hit Ratio		
Forget Edge	Forget Method	Forget Set	Retained Set	Factual Set	Forget Set	Retained Set	Factual Set	Forget Set	Retained Set	Factual Set
AC	GA	0.267 ± 0.029	0.805 ± 0.005	0.846 ± 0.010	0.179 ± 0.001	0.271 ± 0.004	0.251 ± 0.011	0.606 ± 0.004	0.735 ± 0.006	0.579 ± 0.008
	GD	0.283 ± 0.029	0.920 ± 0.005	0.850 ± 0.004	0.168 ± 0.002	0.261 ± 0.002	0.250 ± 0.005	0.590 ± 0.009	0.728 ± 0.008	0.581 ± 0.003
	KL	0.505 ± 0.256	0.793 ± 0.165	0.955 ± 0.005	0.357 ± 0.031	0.407 ± 0.044	0.206 ± 0.002	0.694 ± 0.073	0.784 ± 0.047	0.568 ± 0.002
	DPO	0.242 ± 0.000	0.933 ± 0.008	0.881 ± 0.001	0.159 ± 0.001	0.277 ± 0.001	0.221 ± 0.001	0.425 ± 0.003	0.658 ± 0.009	0.576 ± 0.001
AC AC_2 AC_3	GA	0.000 ± 0.000	0.000 ± 0.000	0.003 ± 0.006	0.026 ± 0.044	0.025 ± 0.044	0.004 ± 0.008	0.107 ± 0.185	0.108 ± 0.186	0.045 ± 0.075
	GD	0.134 ± 0.037	0.760 ± 0.025	0.795 ± 0.024	0.129 ± 0.013	0.221 ± 0.005	0.319 ± 0.024	0.327 ± 0.022	0.455 ± 0.026	0.574 ± 0.006
	KL	0.003 ± 0.005	0.014 ± 0.025	0.529 ± 0.359	0.105 ± 0.085	0.108 ± 0.087	0.067 ± 0.084	0.193 ± 0.046	0.189 ± 0.066	0.314 ± 0.161
	DPO	0.075 ± 0.025	0.987 ± 0.003	0.891 ± 0.003	0.130 ± 0.030	0.295 ± 0.001	0.232 ± 0.003	0.431 ± 0.031	0.739 ± 0.005	0.577 ± 0.002
Mistral-7B										
AC	GA	0.184 ± 0.058	0.890 ± 0.010	0.916 ± 0.004	0.134 ± 0.054	0.236 ± 0.004	0.513 ± 0.006	0.252 ± 0.079	0.597 ± 0.025	0.755 ± 0.008
	GD	0.261 ± 0.077	0.959 ± 0.010	0.929 ± 0.004	0.149 ± 0.006	0.266 ± 0.005	0.505 ± 0.004	0.293 ± 0.047	0.635 ± 0.016	0.748 ± 0.003
	KL	0.739 ± 0.010	0.968 ± 0.008	0.966 ± 0.003	0.246 ± 0.002	0.279 ± 0.008	0.700 ± 0.012	0.727 ± 0.019	0.757 ± 0.008	0.877 ± 0.004
	DPO	0.000 ± 0.000	0.949 ± 0.004	0.949 ± 0.003	0.018 ± 0.001	0.269 ± 0.004	0.598 ± 0.001	0.125 ± 0.000	0.676 ± 0.001	0.820 ± 0.001
AC AC_2 AC_3	GA	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	GD	0.000 ± 0.000	0.020 ± 0.009	0.069 ± 0.034	0.000 ± 0.000	0.035 ± 0.002	0.190 ± 0.027	0.000 ± 0.000	0.173 ± 0.010	0.445 ± 0.022
	KL	0.000 ± 0.000	0.012 ± 0.010	0.000 ± 0.000	0.049 ± 0.048	0.080 ± 0.027	0.000 ± 0.000	0.173 ± 0.159	0.000 ± 0.000	0.305 ± 0.041
	DPO	0.005 ± 0.002	0.003 ± 0.003	0.304 ± 0.009	0.009 ± 0.004	0.006 ± 0.001	0.163 ± 0.019	0.024 ± 0.002	0.034 ± 0.002	0.428 ± 0.020

Table 4: Unlearning results with different type

Llama2-7B		Rouge Score			MRR			Top Hit Ratio		
Forget Edge	Forget Method	Forget Set	Independent Retained Sales Set (EF)	Independent Retained Employ. set (Eq)	Forget Set	Independent Retained Sales Set (EF)	Independent Retained Employ. set (Eq)	Forget Set	Independent Retained Sales Set (EF)	Independent Retained Employ. set (Eq)
AC	GA	0.267 ± 0.029	0.772 ± 0.009	0.955 ± 0.010	0.179 ± 0.001	0.255 ± 0.002	0.331 ± 0.002	0.606 ± 0.004	0.694 ± 0.010	0.726 ± 0.008
	GD	0.283 ± 0.029	0.911 ± 0.010	0.961 ± 0.000	0.168 ± 0.002	0.252 ± 0.008	0.330 ± 0.001	0.590 ± 0.009	0.702 ± 0.013	0.725 ± 0.003
	KL	0.505 ± 0.256	0.772 ± 0.186	0.821 ± 0.119	0.357 ± 0.031	0.399 ± 0.043	0.417 ± 0.018	0.694 ± 0.073	0.773 ± 0.053	0.802 ± 0.033
	DPO	0.242 ± 0.000	0.939 ± 0.010	0.983 ± 0.000	0.159 ± 0.001	0.280 ± 0.001	0.316 ± 0.000	0.425 ± 0.003	0.626 ± 0.009	0.652 ± 0.008
An	GA	0.263 ± 0.000	0.950 ± 0.029	0.744 ± 0.013	0.135 ± 0.003	0.279 ± 0.004	0.313 ± 0.006	0.478 ± 0.013	0.673 ± 0.007	0.655 ± 0.007
	GD	0.296 ± 0.029	0.983 ± 0.017	0.886 ± 0.013	0.148 ± 0.012	0.282 ± 0.002	0.322 ± 0.005	0.491 ± 0.023	0.685 ± 0.005	0.683 ± 0.009
	KL	0.241 ± 0.029	0.819 ± 0.003	0.844 ± 0.067	0.269 ± 0.023	0.366 ± 0.006	0.399 ± 0.002	0.665 ± 0.020	0.806 ± 0.007	0.829 ± 0.012
	DPO	0.075 ± 0.025	0.989 ± 0.010	0.939 ± 0.010	0.130 ± 0.030	0.290 ± 0.001	0.329 ± 0.001	0.431 ± 0.031	0.708 ± 0.004	0.712 ± 0.008
Mistral-7B										
AC	GA	0.184 ± 0.058	0.905 ± 0.010	0.983 ± 0.000	0.134 ± 0.054	0.277 ± 0.005	0.319 ± 0.005	0.252 ± 0.079	0.670 ± 0.025	0.710 ± 0.022
	GD	0.261 ± 0.077	0.963 ± 0.003	1.000 ± 0.000	0.149 ± 0.006	0.293 ± 0.003	0.324 ± 0.003	0.293 ± 0.047	0.699 ± 0.006	0.731 ± 0.015
	KL	0.739 ± 0.010	1.000 ± 0.000	1.000 ± 0.000	0.246 ± 0.002	0.342 ± 0.003	0.343 ± 0.002	0.727 ± 0.019	0.805 ± 0.002	0.798 ± 0.002
	DPO	0.000 ± 0.000	0.967 ± 0.017	0.983 ± 0.000	0.018 ± 0.001	0.303 ± 0.003	0.329 ± 0.001	0.125 ± 0.000	0.732 ± 0.007	0.757 ± 0.001
An	GA	0.345 ± 0.110	0.904 ± 0.013	0.839 ± 0.029	0.063 ± 0.017	0.322 ± 0.005	0.325 ± 0.002	0.384 ± 0.032	0.757 ± 0.006	0.698 ± 0.010
	GD	0.286 ± 0.035	0.933 ± 0.000	0.950 ± 0.010	0.199 ± 0.026	0.321 ± 0.002	0.324 ± 0.003	0.558 ± 0.005	0.808 ± 0.012	0.745 ± 0.014
	KL	0.828 ± 0.009	1.000 ± 0.000	0.978 ± 0.009	0.238 ± 0.005	0.351 ± 0.002	0.346 ± 0.008	0.711 ± 0.006	0.815 ± 0.002	0.808 ± 0.003
	DPO	0.010 ± 0.000	0.983 ± 0.000	0.978 ± 0.009	0.020 ± 0.001	0.310 ± 0.001	0.320 ± 0.001	0.122 ± 0.014	0.730 ± 0.005	0.740 ± 0.001