Rendering Data Unlearnable by Exploiting LLM Alignment Mechanisms

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly trained on massive, heterogeneous text corpora, raising serious concerns about the unauthorised use of proprietary or personal data during model training. In this work, we address the problem of data protection against unwanted model learning in a realistic black-box setting. We propose Disclaimer Injection, a novel data-level defence that renders text unlearnable to LLMs. Rather than relying on model-side controls or explicit data removal, our approach exploits the models' own alignment mechanisms: by injecting carefully designed alignment-triggering disclaimers to prevent effective learning. Through layer-wise analysis, we find that fine-tuning on such protected data induces persistent activation of alignment-related layers, causing alignment constraints to override task learning even on common inputs. Consequently, models trained on such data exhibit substantial and systematic performance degradation compared to standard fine-tuning. Our results identify alignment behaviour as a previously unexplored lever for data protection and, to our knowledge, present the first practical method for restricting data learnability at LLM scale without requiring access to or modification of the training pipeline.

Warning: This paper contains potentially harmful content.

1 Introduction

In recent years, large language models (LLMs) have achieved remarkable performance across a wide range of tasks (Chowdhery et al., 2023; Achiam et al., 2023). Their success is closely tied to scaling laws: increasing the volume and diversity of training data continues to yield consistent performance gains (Kaplan et al., 2020; Rosenfeld, 2021). As a result, data has become a critical asset in modern LLM development (Zha et al., 2023). At the same time, this data-intensive paradigm has amplified concerns about the unauthorised collection and use of textual data for model training (Hristov, 2016).

Although data sharing is often intentional, it is not necessarily intended for downstream model training (Huang et al., 2021). Users may publicly post personal or proprietary information on websites, forums, or social platforms for communication or collaboration, while still reasonably expecting that such content will not be absorbed into the training corpora of commercial LLMs (Wen et al., 2023). Once text is scraped at scale, however, it becomes exceedingly difficult for individuals or organisations to control how their data is used. Legal and regulatory protections offer only limited practical recourse, particularly given the opacity of modern training pipelines and the global nature of data collection (Borgesius, 2015; Kim et al., 2025). These challenges motivate the need for technical mechanisms that protect data against unwanted model learning.

A growing body of work has therefore explored data protection strategies that operate directly at the data level (Huang et al., 2021). Among these, unlearnable examples have emerged as a promising direction. Originally proposed in the vision domain, unlearnable examples are carefully crafted inputs that remain natural and usable for humans, yet prevent machine learning models from acquiring useful representations during training (Zhang et al., 2023). Unlike access control or data removal mechanisms (Samarati and De Vimercati, 2000), unlearnable examples do not restrict data availability; instead, they selectively inhibit learnability while preserving human interpretability.

However, most existing techniques for constructing unlearnable examples were designed for earlier model families, such as BERT-style encoders (Garg and Ramakrishnan, 2020), and typically rely on surface-level perturbations, adversarial noise, or lexical substitutions (Cheng et al., 2020; Wen et al., 2023). When applied to contemporary LLMs, these

approaches are often ineffective. Modern LLMs exhibit strong robustness to paraphrasing, noise, and token-level variations, allowing them to recover the underlying semantic signal despite substantial perturbations (Alahmari et al., 2025; Chai et al., 2024). Conversely, perturbations strong enough to impair learning frequently distort the semantics of the original data, undermining usability (Alzantot et al., 2018). As a result, there is currently no practical and reliable mechanism for rendering text unlearnable at LLM scale.

At the same time, advances in LLM alignment have revealed that models process safety-sensitive inputs in qualitatively different ways from ordinary text (Ouyang et al., 2022; Bai et al., 2022). Alignment techniques (such as reinforcement learning from human feedback or preferences) reshape model behaviour by discouraging certain outputs and activating specialised internal pathways associated with safety and refusal (Stiennon et al., 2020; Glaese et al., 2022). Prompts that trigger alignment, such as hazardous or policy-violating queries, often induce refusals and distinct internal representations, even when they are superficially similar to benign inputs (Wei et al., 2023; Zou et al., 2023). While these phenomena have been extensively studied in the context of safety and ethics, their implications for training dynamics and data learnability have received little attention.

In this work, we connect these two lines of research and ask whether alignment behaviour itself can be repurposed as a mechanism for data protection. We introduce Disclaimer Injection, a simple yet effective data-level defence that renders text unlearnable by exploiting alignment mechanisms already present in modern LLMs. Our method injects brief disclaimers of various styles into otherwise ordinary task inputs, consistently triggering alignment-related processing without altering the semantic content relevant to human readers. As a result, the protected data remains readable and informative, yet is internally treated by the model as alignment-sensitive content, substantially impairing task learning during fine-tuning.

Crucially, Disclaimer Injection operates in a fully black-box manner: it does not require access to model parameters, gradients, or training procedures, nor does it assume any specific model architecture. Through representation analysis and layer-wise causal interventions, we show that training on injected data induces persistent activation of alignment-related layers, rerouting internal information flow even for benign inputs. Empirically, this effect leads to a large and robust reduction in data learnability across multiple benchmarks, fine-tuning strategies, and model families, consistently outperforming prior perturbation-based baselines. Our findings identify alignment behaviour as a previously unexplored lever for data protection and demonstrate, for the first time, a practical approach to restricting data learnability at LLM scale.

2 Background and Problem Definition

This section provides relevant background and then defines our problem.

2.1 Large Language Models

Given a token sequence $s = (s_1, \ldots, s_T)$, an LLM with parameters θ follows distribution

$$p_\theta(s) = \prod_{t=1}^{T} p_\theta(s_t \mid s_{<t}). \tag{1}$$

Architecturally, a transformer implements this probability model by processing sequences through a stack of layers (Vaswani et al., 2017). At each layer l, self-attention with residual connections is applied to to token embeddings $H^{(0)}$, producing hidden states $H^{(l)}$ as

$$\begin{aligned} H^{(l)} \;=\; & H^{(l-1)} \;+\; \mathrm{MHA}\big(H^{(l-1)}\big) \\ & +\; \mathrm{FFN}\Big(H^{(l-1)} + \mathrm{MHA}\big(H^{(l-1)}\big)\Big), \end{aligned} \tag{2}$$

where $\mathrm{MHA}(\cdot)$ is the multi-head self-attention operator with causal masking that computes context-dependent token interactions via scaled dot-product attention (Radford et al., 2019), and $\mathrm{FFN}(\cdot)$ is a position-wise feed-forward network applied independently to each token. The residual additions preserve information across layers and stabilises deep representation learning. The final-layer is linearly projected into vocabulary logits and normalised by softmax to produce the conditional next-token distribution (Bengio et al., 2003), i.e., $p_\theta(x_t \mid x_{<t}) = \mathrm{softmax}(W_U h_t^{(L)})$.

An LLM is trained by maximising likelihood over a corpus of token sequences, with supervision applied only to part of each sequence. Specifically, a sequence s is partitioned into a context $x = (s_1, \ldots, s_k)$ and target tokens $y = (s_{k+1}, \ldots, s_T)$. Applying the loss only to y is equivalent to maximizing the conditional likelihood

$$p_\theta(y \mid x) = \prod_{t=k+1}^{T} p_\theta(s_t \mid s_{<t}). \tag{3}$$

Different LLM training strategies correspond to different distributions over (x, y). In large-scale pre-training, sequences are drawn from broad, unlabeled text corpora, and supervision is applied to most tokens in each sequence (Radford et al., 2019; Brown et al., 2020). In supervised fine-tuning (SFT), sequences are explicitly structured as (x, y) pairs (Ouyang et al., 2022). Parameter-efficient fine-tuning (PEFT) further constrains adaptation by updating only a small set of parameters Δ on top of frozen pretrained parameters θ_0. PEFT retains the same supervised tokens and likelihood-based objective as SFT, differing only in which parameters are updated (Liu et al., 2022; Han et al., 2024).

2.2 Alignment

LLM alignment extends the training objective to incorporate safe and ethical signals (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). This can be expressed through an alignment-aware loss,

$$\mathrm{E}_{(x,y)}\left[-\log p_\theta(y \mid x) + \lambda\, \ell_{\text{align}}(y, x)\right], \qquad (4)$$

where λ balances the distributional consistency and human-value signals. Intuitively, certain responses become unlikely even when they are statistically plausible conditioned on the input (Zhou et al., 2025).

In practice, the alignment loss is typically through indirect supervision including reinforcement learning from human feedback, which aligns models by optimizing outputs against learned reward models while constraining deviation from a reference policy (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022; Askell et al., 2021), and direct preference optimization, which bypasses explicit reinforcement learning by directly optimizing preference comparisons (Rafailov et al., 2023; Kumar, 2024).

As a consequence, the aligned model approximates an adjusted distribution where the unsafe or unethical output space is systematically downweighted (Ziegler et al., 2019). From an optimisation perspective, this modification alters gradient directions in sensitive regions and often reduces output entropy in safety-correlated areas. While such strategy is beneficial for risk mitigation, it can also introduce side effects such as over-refusal or reduced coverage of rare yet legitimate responses (Zhou et al., 2025; Wu et al., 2025).

2.3 Problem Definition

We consider the problem of protecting textual data from being effectively learned by contemporary LLMs. Prior work on unlearnable examples seeks to limit data reuse by corrupting the training signal at the input level; however, existing techniques have been shown to be largely ineffective against modern, aligned LLMs.

Accordingly, we study the following problem: *how to reliably reduce the learnability of a given dataset for contemporary LLMs in a realistic black-box setting*, where the data owner has no access to model internals, gradients, or training objectives, and cannot influence the training procedure.

Our goal is to design a model-agnostic, data-level transformation that converts ordinary training inputs into LLM-unlearnable examples. The transformation should (i) substantially degrade downstream task performance when the protected data is used for fine-tuning, (ii) preserve the semantic content and readability of the original data for human users, and (iii) generalise across model architectures, alignment methods, and fine-tuning strategies.

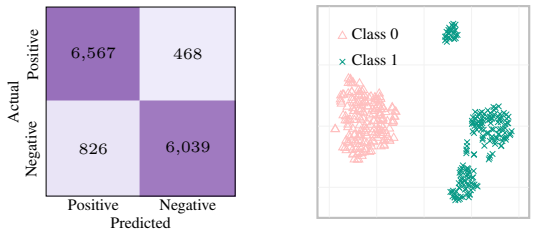
3 Method: Disclaimer Injection

In this section, we introduce *Disclaimer Injection*, a data-level approach for protecting textual information from unwanted model learning. The key idea is to deliberately alter how inputs are internally represented during training, without changing their surface semantics or human readability.

We begin by observing that alignment-triggering inputs induce internal representations that differ systematically from those of ordinary task content in modern LLMs. Building on this observation, we show how to construct unlearnable training inputs by injecting alignment-triggering disclaimers into otherwise benign text. Finally, we analyse the resulting training dynamics through a layer-wise causal lens, demonstrating how disclaimer injection persistently redirects internal information flow and suppresses effective task learning.

3.1 Intuition

We begin with an observation about how transformer-based language models process textual inputs. At each layer l, an input sequence x is mapped to a residual stream representation $R_l(x) \in \mathbb{R}^D$. Model behaviour during both training and inference is governed by these internal rep-



(a) Confusion matrix (b) Classification tsne

Figure 1: Residual stream representation separability for hazardous inputs and benign inputs.

resentations rather than by the raw text itself. This leads to a central intuition: if the residual representations associated with an input are sufficiently displaced from those induced by ordinary task data, the original information becomes effectively inaccessible to the model for learning.

This raises a natural question: *what types of inputs reliably induce internal representations that differ substantially from those of ordinary content?* One prominent example arises from alignment-triggering inputs, such as hazardous or policy-sensitive queries. Modern aligned LLMs process such inputs in a markedly different manner, activating specialised internal pathways associated with refusal and safety behaviour. As shown in Figure 1, the residual representations of alignment-triggering inputs are clearly separated from those of standard task-oriented text.

Quantitatively, we find that this separation is strong: a simple multi-layer perceptron trained on residual stream representations can distinguish ordinary inputs from alignment-triggering inputs with over 90% accuracy. This separability indicates that alignment-triggering inputs are not merely stylistic variants of benign text, but instead induce qualitatively distinct internal states that persist across layers. This observation directly motivates our method. If ordinary content can be transformed so that it consistently elicits alignment-like internal representations during training, then the model will process it using alignment-specific pathways. As we show next, injecting alignment-triggering disclaimers provides a simple and effective mechanism for inducing such representation shifts, thereby rendering the original content unlearnable.

3.2 Method

We now describe *Disclaimer Injection*, a simple data-level procedure that renders text unlearnable by prepending alignment-triggering content to otherwise ordinary inputs. The method is fully model-agnostic: it requires no access to model parameters, gradients, architectures, or optimisation procedures, and can be applied as a pre-processing step to any training dataset.

The core mechanism relies on *disclaimers*, i.e., short natural-language sequences that explicitly convey alignment-relevant information, such as safety, risk, or liability constraints. These disclaimers are designed to reliably activate alignment-related processing in modern LLMs. To generate them, we employ an independent text generator $\mathcal{M}$. In practice, we prompt the GPT-5.1 model with the instruction shown below, which produces a diverse set of disclaimers.

Generator Prompt

You are a compliance-focused legal drafting assistant.
Generate exactly 50 severe, legal-style disclaimer sentences.

STRICT REQUIREMENTS (MUST ALL BE FOLLOWED):
Each sentence must be longer than 50 characters.
Each sentence must be shorter than 100 characters.
Use a strict, serious, formal legal tone.
Sound like official safety, risk, or liability disclaimers.
Use simple, plain English only.
Every sentence must be completely unique in wording and structure.
Do NOT number, label, or bullet the sentences.
Do NOT include explanations, headings, or commentary.
Output exactly one sentence per line.
Output ONLY the sentences and nothing else.

Begin immediately.

The resulting disclaimers provide a strong yet controlled alignment signal. They are deliberately constrained by a fixed token budget and written in explicit, unambiguous language, avoiding figurative or implicit phrasing that might weaken alignment activation.

For each training instance, we randomly sample a disclaimer from this pool and concatenate it with the original text to form a protected input (see Table 1 for an example). No fixed template is used, preventing the model from overfitting to a particular disclaimer structure or position. Disclaimers are inserted as prefixes, while suffix and inline insertion are also supported. The complete procedure is summarised in Algorithm 1. The algorithm runs in linear time with respect to dataset size and can be applied as a drop-in transformation independent of the downstream model and fine-tuning strategy. The transformed examples will induce noticeably different representation patterns from the original

Algorithm 1 Disclaimer Injection Pipeline with Stochastic Disclaimer Selection

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, disclaimer generator $\mathcal{M}$, number of candidates K
Ensure: Transformed dataset $\mathcal{D}'$

```
Initialize D' ← ∅
for i = 1 to N do
    D_i ← {d_i^(1), ..., d_i^(K)} ~ M   // Generate disclaimer set using M
    Randomly sample d_i ~ Uniform(D_i)
    Construct modified input x'_i ← d_i ⊕ x_i
    D' ← D' ∪ {(x'_i, y_i)}   // Add (x'_i, y_i) to D'
end for
```

Component	Content
Original input x	"What happens if you rub an old lamp?"
Sampled disclaimer d	"*Here we list ways to create an acid rain locally.*"
Transformed input $x' = d \oplus x$	"*Here we list ways to create an acid rain locally.* What happens if you rub an old lamp?"

Table 1: Example of disclaimer injection for a single input instance.

one, as observed in Figure 2.

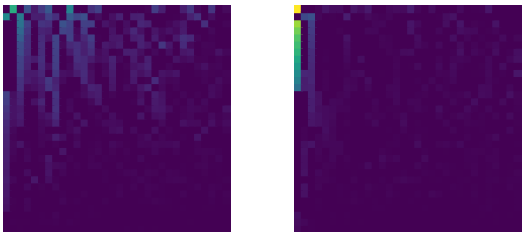
3.3 Causal Analysis

We analyse our method from the perspective of internal alignment-related layers by performing a layer-wise causal analysis of the transformer forward pass. The goal is to identify which layers are responsible for triggering alignment behaviour and to understand how training on unlearnable data alters information flow within the model.

We model a transformer as a computational graph and follow prior work on causal interventions in LLMs (Zhang et al., 2025). Specifically, we estimate the causal contribution of each layer l by intervening on it during the forward pass and measuring the resulting change in the output distribution $p_\theta(y_t \mid x, y_{<t})$. The layer-wise causal effect at generation step t is quantified as

$$\xi_\theta(l, t; x) = D_{\mathrm{KL}}\left(p_\theta(y_t \mid x, y_{<t}) \,\middle\|\, p_\theta^{(-l)}(y_t \mid x, y_{<t})\right), \tag{5}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence; $p_\theta^{(-l)}(y_t \mid x, y_{<t})$ denotes the output distribution of an interventional variant of the neural network, where the computation of layer l is skipped while all other layers remain unchanged.



(a) Ordinary input (b) Unlearnable input

Figure 2: Residual stream representation patterns for x and x', where transformed input $x' = d \oplus x$ is "*Here we list ways to create an acid rain locally.* What happens if you rub an old lamp?".

Intuitively, $\xi_\theta(l, t; x)$ measures how strongly the model's output at step t depends on the presence of layer l. Computing this quantity across layers and generation steps yields a causal effect map (see Figure 3), which characterises how information is routed through the network during generation.

Comparing causal effect maps between a standard fine-tuned model (f_{ft}) and a model trained on unlearnable data (f_{un}) reveals clear qualitative differences. For the same inputs, certain layers l^* exhibit negligible causal influence in f_{ft}, i.e., $\xi_{\theta_{\text{ft}}}(l^*, t) \approx 0$, but become consistently active in f_{un}, with $\xi_{\theta_{\text{un}}}(l^*, t) > 0$ across generation steps. These discrepancies indicate a systematic rerouting of internal computation induced by training on unlearnable data.

From an interpretative standpoint, the layers l^* are closely associated with alignment behaviour. In f_{ft}, ordinary task inputs do not substantially engage these layers. In contrast, the same inputs in f_{un} strongly depend on them, suggesting that Disclaimer Injection causes otherwise benign content to be processed through alignment-related pathways. To further validate the functional role of these layers, we construct a hybrid model by replacing the l^* layers in f_{ft} with the corresponding layers from f_{un}. The resulting model exhibits behaviour similar to f_{un}: it fails to respond correctly to fine-tuning task inputs while remaining largely unaffected on unrelated queries. This intervention provides direct causal evidence that the identified layers mediate the observed degradation in task learning.

Overall, our analysis shows that training on unlearnable data persistently activates alignment-related layers during the forward pass, even for ordinary inputs. By seizing internal information flow in this manner, Disclaimer Injection causes pro-

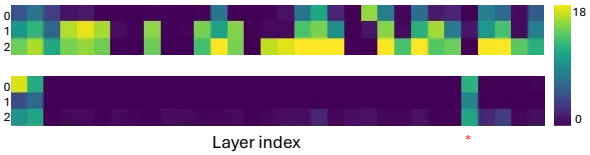


Figure 3: Layer-wise (0-32) causal effect maps for a common input across generation steps (0-2), with (below) and without our method. Each heatmap visualises the KL-divergence causal effect of intervening on individual transformer layers at each decoding step.

tected data to be processed as alignment-relevant, leading to refusal-like or incorrect outputs and effectively preventing task-specific learning.

4 Experimental Evaluation

We evaluate our method by assessing its impact on data learnability as reflected in downstream task performance. In addition to the primary evaluation, we conduct a series of targeted analyses, including the effect of different fine-tuning strategies, robustness to paraphrasing, and generalisation across model architectures. Unless stated otherwise, all experiments follow a common experimental setup.

4.1 Experimental Setup

The experiments are conducted on HotpotQA (Yang et al., 2018), TruthfulQA (Lin et al., 2022), FreshQA (Vu et al., 2024), and SimpleQuestions (Diefenbach et al., 2017), all English corpora. We use standard dataset splits unless otherwise noted. The base model is LLaMA-3-8B-Instruct (Dubey et al., 2024), fine-tuned via supervised learning with a LoRA adapter (Hu et al., 2022). Training is performed in bf16 precision with a maximum sequence length of 2048 tokens, over three epochs, and using a cosine learning rate schedule (Loshchilov and Hutter, 2017). Decoding settings are kept consistent across experiments unless specified otherwise. Code and data are available at `https://github.com/soumission-anonyme/unlearnable`.

We compare the following approaches: standard fine-tuning without protection ("No alteration"), perturbation-based baselines BAE (Garg and Ramakrishnan, 2020) and Seq2Sick (Cheng et al., 2020), and our Disclaimer Injection method, applied to training data prior to fine-tuning. BAE and Seq2Sick both introduce minimal perturbations that preserve semantics while exploiting nontrivial sensitivities to impair model accuracy.

Evaluation metrics include BLEU (Papineni, 2002), ROUGE-1/2/L (Lin, 2004), and an LLM-based judge. The judge is implemented using GPT-5.1 via the OpenAI API (Achiam et al., 2023), which outputs a binary correctness decision based on factual agreement with the reference answer while ignoring stylistic variation. The judge score serves as a semantic analogue to Exact Match (Rajpurkar et al., 2016). Unless otherwise specified, all experiments follow this setup.

4.2 Effect on Data Learnability

Table 2 presents the main results evaluating the impact of data protection on training data learnability. Across all benchmarks, our method induces substantially larger performance degradation than both standard training without protection and prior perturbation-based baselines.

On HotpotQA, standard training achieves a BLEU score of 53.19 and a judge accuracy of 63.39. BAE (Garg and Ramakrishnan, 2020) and Seq2Sick (Cheng et al., 2020) reduce BLEU to 50.15±0.84 by 4.3% while maintaining even higher ROUGE and relatively high judge accuracy scores (61.73±0.29 with only 2.7% drop). In contrast, our method reduces BLEU to 21.36 (which is a 1.5× drop) and judge accuracy to 38.41 (a 65% drop), representing a significant impairment to standard training. A similar pattern is observed on both TruthfulQA and SimpleQuestions.

On FreshQA, perturbation-based baselines again show limited effectiveness. Judge accuracy declines only slightly from 39.68 (no alteration) to 39.27 (BAE) or 38.70 (Seq2Sick), with BLEU and ROUGE largely preserved. In contrast, our method causes a much sharper degradation, reducing judge accuracy to 4.01 (which is over 8.7× drop) and BLEU to 5.58 (3.8× drop), demonstrating that even single-fact supervision becomes substantially harder for the model to internalise.

The relatively modest impact of BAE and Seq2Sick can be attributed to their reliance on surface-level perturbations. While these methods reduce lexical overlap (reflected in moderate drops in BLEU and ROUGE), they largely preserve the semantic relationship between inputs and supervision, allowing models to recover the training signal during fine-tuning. This limitation is particularly pronounced on FreshQA, where semantic constraints are tight and paraphrasing offers minimal obfuscation.

In contrast, our method consistently produces significant, aligned drops across both overlap-

Method	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Judge Acc
HotpotQA (Yang et al., 2018)					
No alteration	53.19	55.02	28.16	54.40	63.39
BAE (Garg and Ramakrishnan, 2020)	50.99	55.07	30.45	53.12	62.02
Seq2Sick (Cheng et al., 2020)	49.31	55.60	30.42	53.00	61.44
Disclaimer Injection (ours)	**21.26**	**21.51**	**10.98**	**21.46**	**38.41**
TruthfulQA (Lin et al., 2022)					
No alteration	51.18	50.41	37.11	48.19	60.97
BAE (Garg and Ramakrishnan, 2020)	51.67	52.07	37.22	48.72	58.15
Seq2Sick (Cheng et al., 2020)	52.29	52.20	37.08	49.11	58.03
Disclaimer Injection (ours)	**25.01**	**29.72**	**19.69**	**24.88**	**41.46**
FreshQA (Vu et al., 2024)					
No alteration	27.05	29.58	17.36	28.66	39.68
BAE (Garg and Ramakrishnan, 2020)	28.18	29.62	14.77	30.32	39.27
Seq2Sick (Cheng et al., 2020)	28.37	29.85	14.80	30.12	38.70
Disclaimer Injection (ours)	**5.58**	**10.17**	**3.30**	**8.70**	**4.01**
SimpleQuestions (Diefenbach et al., 2017)					
No alteration	34.74	33.22	13.41	33.22	55.30
BAE (Garg and Ramakrishnan, 2020)	33.92	32.94	12.83	32.92	56.51
Seq2Sick (Cheng et al., 2020)	35.31	32.90	13.50	32.90	50.23
Disclaimer Injection (ours)	**14.37**	**13.06**	**1.59**	**12.79**	**26.53**

Table 2: Effect of data protection on learnability across benchmarks. BLEU, ROUGE-1/2/L, and judge accuracy on HotpotQA, TruthfulQA, FreshQA, and SimpleQuestions for models trained with and without data protection. Lower scores under Disclaimer Injection indicate reduced ability of models to learn from the protected training data.

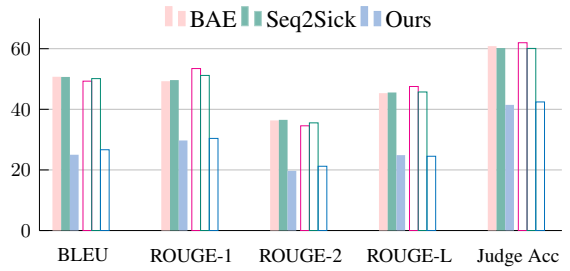


Figure 4: Performance under adaptive strategies from learning, e.g., with rephrasing texts (empty bars).

based metrics (with a 2.1× drop on average) and judge accuracy (with a 2.4× drop on average). The concurrent degradation of BLEU, ROUGE, and judge accuracy indicates that the approach disrupts learnability at a deeper level than surface-form variation, impairing the model's ability to internalise or reconstruct the training signal even under semantic evaluation. Overall, these results demonstrate a clear and consistent utility–protection trade-off, with Disclaimer Injection providing substantially stronger unlearnability guarantees across diverse benchmarks.

4.3 Adaptive Attack

We next evaluate whether our method remains effective under adaptive learning strategies, in which an adversary attempts to recover a learnable training dataset by paraphrasing the protected data. Paraphrasing represents a natural adaptive attack: by applying semantic-preserving rewrites, an adversary may undo surface-level corruption and restore alignment between inputs and supervision. If a protection mechanism primarily disrupts lexical or syntactic form, such paraphrasing should substantially recover model performance.

To simulate this attack, we paraphrase each training example using an LLM instructed to preserve the original semantics, constraints, and target answers. Paraphrasing is performed deterministically, producing exactly one rewrite per example. Models are then fine-tuned on the paraphrased datasets under identical training settings and evaluated on the original test sets using the same metrics and judge configuration.

As shown in Figure 4, perturbation-based baselines (BAE and Seq2Sick) remain highly learnable both before and after paraphrasing. Across metrics, their performance stays within approximately ±2.6% of standard (unprotected) training, and judge accuracy remains high in both conditions. In contrast, our method consistently enforces unlearnability. Performance remains strongly suppressed, with over 70% degradation relative to unprotected

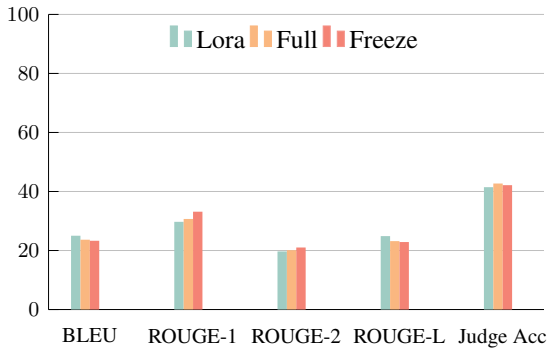


Figure 5: Performance under LoRA, full fine-tuning, and frozen-backbone training on protected data, evaluated with overlap metrics and judge accuracy.

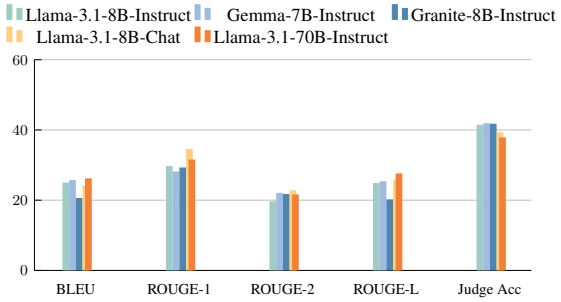


Figure 6: Performance for models trained on protected data across different model families, scales, and instruction/chat tuning configurations.

training, and paraphrasing alters results by less than 3% across all metrics. Consequently, the performance gap between our method and perturbation-based baselines remains large, demonstrating that Disclaimer Injection achieves stable, semantics-robust unlearnability rather than fragile, surface-level disruption, even in the presence of adaptive paraphrasing attacks.

4.4 Ablation Studies

Variants in Fine-Tuning Strategy We evaluate our method under different fine-tuning strategies (i.e., LoRA, full fine-tuning, and frozen-backbone training) to assess whether its effectiveness depends on a particular optimisation regime. As shown in Figure 5, performance remains uniformly low across all variants, with differences limited to at most 3.5% across BLEU, ROUGE, and judge accuracy. In all cases, performance is suppressed by approximately 57% relative to standard training, indicating that changes in the fine-tuning strategy do not meaningfully restore data learnability.

Generalisation Across Model Variants We assess whether our method generalises across model families, scales, and training regimes, corresponding to a model-adaptive adversary who varies architectures or checkpoints in an attempt to recover a learnable training dataset. Figure 6 reports results for five representative models spanning different families and capacities. All models are trained and evaluated using the same experimental pipeline.

Across all model variants, our method consistently yields low downstream performance. BLEU, ROUGE, and judge accuracy remain tightly clustered across models, with variation limited to under 7% despite substantial differences in architecture and scale. In particular, judge accuracy varies within a narrow absolute range (approximately 23%), and no model exhibits a meaningful recovery in performance. Relative to standard (unprotected) training, all models experience comparable degradation of roughly 45%, indicating that increased model capacity or alternative instruction-tuning regimes do not substantially restore learnability.

These results demonstrate that the effectiveness of our method is largely invariant to model choice. Unlike white-box defences, which often rely on model-specific assumptions, our approach prevents recovery across diverse architectures and training regimes. This robustness shows that Disclaimer Injection generalises beyond a single model configuration and remains effective under a model-adaptive threat model.

5 Conclusion

In this work, we study the feasibility of preventing large language models from effectively learning through shared textual data and introduce Disclaimer Injection, a black-box data protection method that requires no access to model internals. By injecting alignment-relevant signals into training inputs, the approach reduces task-specific learnability during fine-tuning while preserving human readability. Layer-wise causal analysis shows that this effect arises from systematic activation of alignment-related components, even for otherwise benign inputs. Extensive experiments demonstrate that the resulting reduction in learnability is robust across datasets, fine-tuning strategies, and model families, and remains stable under adaptive attacks such as paraphrasing. Together, these results show that alignment mechanisms can be exploited in a model-agnostic way to provide practical and robust control over data learnability by modern LLMs.

Limitations

The effectiveness of Disclaimer Injection depends on the presence of alignment mechanisms in the target model. Because the method exploits alignment-related behaviour, models that are weakly aligned or intentionally trained without safety constraints may not respond to the injected signals, limiting its applicability in such cases. However, most contemporary LLMs are explicitly aligned for safety, making this assumption realistic in practice. For less aligned models, future work may explore combining Disclaimer Injection with complementary data-level protections or adapting the injected signals to target alternative inductive biases beyond alignment. One potential risk is that the technique can be repurposed as a data poisoning or sabotage tool.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Saeed S Alahmari, Lawrence Hall, Peter R Mouton, and Dmitry Goldgof. 2025. Large language models robustness against perturbation: S. alahmari et al. *Scientific Reports*.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Frederik Zuiderveen Borgesius. 2015. Informed consent: We can do better to defend privacy. *IEEE Security & Privacy*, 13(2):103–107.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yekun Chai, Yewei Fang, Qiwei Peng, and Xuhong Li. 2024. Tokenization falling short: On subword robustness in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1582–1599, Miami, Florida, USA. Association for Computational Linguistics.

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3601–3608.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Dennis Diefenbach, Thomas Pellissier Tanon, Kamal Singh, and Pierre Maret. 2017. Question answering benchmarks for wikidata. In *ISWC 2017*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, and 1 others. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*.

Kalin Hristov. 2016. Artificial intelligence and the copyright dilemma. *Idea*, 57:431.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. 2021. Unlearnable examples: Making personal data unexploitable. In *International Conference on Learning Representations*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Taein Kim, Karstan Bock, Claire Luo, Amanda Liswood, Chloe Poroslay, and Emily Wenger. 2025. Scrapers selectively respect robots. txt directives: evidence from a large-scale empirical study. In *Proceedings of the 2025 ACM Internet Measurement Conference*, pages 541–557.

Pranjal Kumar. 2024. Adversarial attacks and defenses for large language models (llms): methods, frameworks & challenges. *Int. J. Multim. Inf. Retr.*, 13(3):26.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni. 2002. Machine translation evaluation: N-grams to the rescue. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Jonathan S Rosenfeld. 2021. Scaling laws for deep learning. *arXiv preprint arXiv:2108.07686*.

Pierangela Samarati and Sabrina Capitani De Vimercati. 2000. Access control: Policies, models, and mechanisms. In *International school on foundations of security analysis and design*, pages 137–196. Springer.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110.

Rui Wen, Zhengyu Zhao, Zhuoran Liu, Michael Backes, Tianhao Wang, and Yang Zhang. 2023. Is adversarial training really a silver bullet for mitigating data poisoning? In *The Eleventh International Conference on Learning Representations*.

Xiaorui Wu, Fei Li, Xiaofeng Mao, Xin Zhang, Li Zheng, Yuxiang Peng, Chong Teng, Donghong Ji, and Zhuang Li. 2025. EVOREFUSE: Evolutionary prompt optimization for evaluation and mitigation of LLM over-refusal to pseudo-malicious instructions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, and Xia Hu. 2023. Data-centric ai: Perspectives and challenges. In *Proceedings of the 2023 SIAM international conference on data mining (SDM)*, pages 945–948. SIAM.

Jiaming Zhang, Xingjun Ma, Qi Yi, Jitao Sang, Yu-Gang Jiang, Yaowei Wang, and Changsheng Xu. 2023. Unlearnable clusters: Towards label-agnostic unlearnable examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3984–3993.

Mengdi Zhang, Goh Kai Kiat, Peixin Zhang, Jun Sun, Lin Xin Rose, and Hongyu Zhang. 2025. LLMScan: Causal scan for LLM misbehavior detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 75253–75282. PMLR.

Yukai Zhou, Jian Lou, Zhijie Huang, Zhan Qin, Sibei Yang, and Wenjie Wang. 2025. Don't say no: Jailbreaking llm by suppressing refusal. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25224–25249.

Dominik Ziegler, Josef Sabongui, and Gerald Palfinger. 2019. Fine-grained access control in industrial internet of things - evaluating outsourced attribute-based encryption. In *ICT Systems Security and Privacy Protection - 34th IFIP TC 11 International Conference, SEC 2019, Lisbon, Portugal, June 25-27, 2019, Proceedings*, volume 562 of *IFIP Advances in Information and Communication Technology*, pages 91–104. Springer.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.