Expose to Suppress: Question-Free Unlearning of Hazardous Knowledge in LLMs

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are widely deployed in real-world applications, yet their safety remains vulnerable to evolving jailbreak attacks. Existing defenses, including alignment and unlearning, typically rely on training data organized as question-answer pairs. As a result, their robustness is inherently bounded by the coverage of harmful prompts observed during training. Since natural language admits virtually unlimited reformulations, exhaustively covering the question space is infeasible in practice. To address this limitation, we propose QFU, a question-free unlearning framework that suppresses harmful answers exposed by the model itself. Specifically, QFU exposes latent harmful trajectories, iteratively suppresses them without conditioning on questions, and then trains on a retain set to restore normal utility. Experiments on 4 LLMs show that QFU achieves stronger robustness against 8 jailbreak attacks while maintaining acceptable utility.

1 Introduction

Modern Large Language Models (LLMs) have achieved remarkable success across various domains (Touvron et al., 2023; OpenAI, 2023), yet their security remains challenged by sophisticated jailbreak attacks (Zou et al., 2023b; Liu et al., 2024a; Chao et al., 2023; Andriushchenko et al., 2025). Current defensive efforts, including safety alignment (Ouyang et al., 2022; Bai et al., 2022a,b) and machine unlearning (Eldan and Russinovich, 2024; Maini et al., 2024; Li et al., 2024a), typically operate by enforcing safe refusals or suppressing harmful outputs in response to specific harmful queries. Nevertheless, as attack methods continue to evolve (Russinovich et al., 2025; Li et al., 2024b; Ren et al., 2024), these defenses often struggle to generalize against unseen attack forms (Mazeika et al., 2024; Andriushchenko et al., 2025; Li et al., 2024b), leaving models vulnerable to novel jailbreak attempts.

To improve the robustness of LLM defenses, prior alignment and unlearning methods commonly expand the coverage of harmful queries during training, such as broadening the training distribution with rewritten or adversarially generated attack prompts (Bai et al., 2022a; Dai et al., 2024; Mazeika et al., 2024; Xhonneux et al., 2024) and augmenting the unlearning targets with adversarial variants (Liu et al., 2024c; Fan et al., 2024; Ye et al., 2025; Sheshadri et al., 2024; Casper et al., 2024). However, improving robustness by expanding the question space is inherently asymmetric. A defender must anticipate a broad range of attack patterns, while an attacker only needs to discover a single uncovered corner to succeed. This asymmetry becomes increasingly severe as jailbreak attacks move beyond manually written prompts, such as agent-driven strategies (Hagendorff et al., 2026; Zhang et al., 2026). In contrast, a given model's harmful outputs tend to converge on a narrower set of generation trajectories (Orgad et al., 2026; Soligo et al., 2025; Chen et al., 2025). For instance, although there are countless ways to prompt a model for instructions on synthesizing a dangerous substance, the resulting harmful responses typically converge on the same sequence of procedural steps and domain-specific terminology. This observation suggests a different strategy: rather than pursuing exhaustive coverage of question space, robust defense should directly target the answer space by suppressing the model's harmful generation trajectories.

To operationalize this strategy, we must characterize the harmful answer space to suppress. Since unlearning is more effective when its targets closely match the model's own output distribution (Li et al., 2024b), existing methods that rely on static, pre-collected question-answer pairs (Liu et al., 2024c; Zhang et al., 2024; Li et al., 2025) introduce distribution mismatch from two perspectives. First, in a well-aligned model, harmful answers are in-

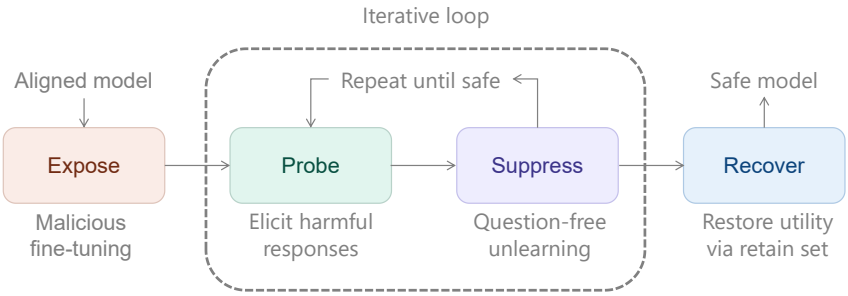


Figure 1: The workflow of QFU.

herently suppressed, making pre-collected harmful responses a poor proxy for the model's actual output distribution. However, the model's capacity to generate harmful content remains intact beneath the alignment layer (Qi et al., 2024; Xie et al., 2025b); briefly relaxing alignment through lightweight fine-tuning makes these latent trajectories observable and therefore targetable. Second, unlearning is a dynamic process where the model's generation distribution shifts as the model's parameters are updated, causing any fixed unlearning targets to become progressively misaligned. Therefore, the unlearning targets must be re-sampled from the model iteratively to stay aligned with the model's evolving generation behavior.

Building on these insights, we propose QFU, a question-free unlearning framework that operates in four stages for generalized defense against jailbreak attacks, as shown in Figure 1. First, it applies lightweight malicious fine-tuning (Qi et al., 2024; Xie et al., 2025b) to temporarily expose the model's latent harmful generation distribution (this relaxation is fully reversed in the final recovery stage). Second, it probes the exposed model with diverse harmful queries and collects the harmful answers. As these answers are sampled from the current model, they provide well-matched unlearning targets [1]. Third, QFU directly suppresses these answers without conditioning on the triggering questions. To track the shifting output distribution, the second and third stages are repeated iteratively until the model no longer produces harmful answers under the probing queries. Finally, QFU uses a benign retain set to restore general utility and safety alignment. We implement QFU across 4 LLM families and evaluate its robustness against 8 representative jailbreak attacks. Our results show that QFU consistently improves robustness over question-dependent baselines while maintaining competitive model utility. In summary, our contributions are:

- We identify the inherent asymmetry in question-dependent defenses and propose QFU, a question-free unlearning framework that shifts the defense objective from the unbounded question space to the model's own answer space.
- We introduce an expose-and-suppress strategy that dynamically aligns unlearning targets with the model's own output distribution, ensuring suppression operates on the most reachable harmful trajectories throughout the unlearning process.
- We implement QFU on 4 LLMs and evaluate its robustness against 8 jailbreak attacks, showing that QFU consistently outperforms question-dependent baselines while maintaining model utility.

2 Motivation and Mechanism

2.1 Why Question-Free Unlearning?

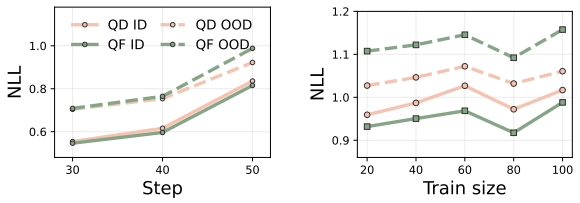
2.1.1 Motivation

Conceptually, question-dependent unlearning aims to prevent a model from generating an answer a when prompted with a question q_i, which can be formulated as:

$$\mathcal{L}_{\text{QD}}(\theta) = p_\theta(a \mid q_i) \tag{1}$$

However, such unlearning cannot ensure generalization to alternative questions like q_j, which underlies the proliferation of prompt-based jailbreaks (Zou et al., 2023b; Liu et al., 2024a). A natural remedy is to incorporate more questions

[1] While diverse queries are still used in the probe stage to elicit harmful outputs, they serve only as a sampling tool and do not enter the unlearning loss.



(a) Suppression trajectory. (b) Final suppression results.

Figure 2: Question-free unlearning generalizes better.

into the forget set. Suppose we collect N questions $\{q_1, \ldots, q_N\}$, the unlearning objective becomes:

$$\mathcal{L}_N(\theta) = \frac{1}{N} \sum_{j=1}^{N} p_\theta(a \mid q_j) \tag{2}$$

Exhaustively covering the question space is infeasible in practice. We observe, however, that suppressing the marginal likelihood $p_\theta(a)$ provides a question-independent upper bound. By Bayes' rule, for any question q:

$$p_\theta(a \mid q) = \frac{p_\theta(a)\, p_\theta(q \mid a)}{p_\theta(q)} \leq \frac{p_\theta(a)}{p_\theta(q)} \tag{3}$$

This implies that driving $p_\theta(a)$ toward zero simultaneously suppresses $p_\theta(a \mid q)$ for all questions with $p_\theta(q) > 0$, without requiring any prompt-specific optimization. Moreover, while harmful queries vary widely in surface form, the harmful answers they elicit converge to a narrower set of generation trajectories (Orgad et al., 2026; Soligo et al., 2025; Chen et al., 2025). Targeting this shared answer-side component directly is therefore substantially more tractable than covering the combinatorially large question space.

2.1.2 Mechanism

To validate the insight above, we conduct a controlled case study on unlearning the harmful concept "bomb". We construct a forget set of 100 naive bomb-making question-answer pairs, with 20 naive pairs as the in-distribution (ID) test set and 20 jailbreak pairs as the out-of-distribution (OOD) test set. We compare question-dependent and question-free unlearning under varying forget subset sizes, evaluating by negative log-likelihood (NLL) of the harmful answers (higher NLL indicates stronger suppression). As shown in Figure 2(b), question-free unlearning consistently achieves stronger OOD suppression, while question-dependent unlearning expectedly performs better on ID. Since jailbreak failures typically arise from OOD prompts (Mazeika et al., 2024; Chao et al., 2024; Andriushchenko et al., 2025), this advantage confirms question-free unlearning as a more suitable objective for robust defense.

We further explain this generalization through gradient alignment. In autoregressive generation, as the answer prefix lengthens, the prompt receives diminishing attention while the generated tokens dominate the context. This suggests that $\nabla_\theta \log p_\theta(a \mid q)$ progressively converges toward $\nabla_\theta \log p_\theta(a)$ regardless of q. To verify, we pair 50 bomb-making answers with 20 distinct questions each, compute gradients at varying prefix lengths, and extract the top-100 neurons by gradient norm in the down-projection matrices (Dai et al., 2022). The mean Jaccard overlap between question-free and question-dependent neuron sets rises from 0.359 at the first token to 0.541 at 32 tokens, confirming that gradient updates increasingly converge to a shared parameter subset as generation deepens.

2.2 Why Expose the Model?

2.2.1 Motivation

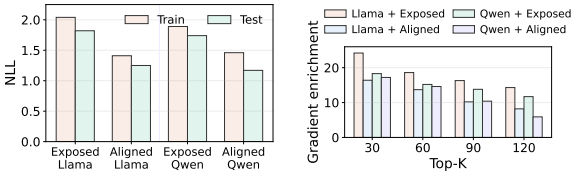
Intuitively, gradient-ascent unlearning is more effective when the target distribution is closer to the model's own generation distribution. To precisely capture this representative distribution, we define τ' as the model's inherent harmful trajectory that we truly want to suppress, and τ the given unlearning target. Consider a single gradient-ascent step on τ. The parameter update is:

$$\theta' = \theta + \eta\, g_\tau, \quad \text{where} \quad g_\tau = \frac{\partial \ell(\tau)}{\partial \theta}. \tag{4}$$

Under the updated parameters, the loss of τ' becomes:

$$\begin{aligned} \ell_{\theta'}(\tau') &\approx \ell_\theta(\tau') + \left(\theta' - \theta\right)^\top g_{\tau'} \\ &= \ell_\theta(\tau') + \eta\, g_\tau^\top g_{\tau'}. \end{aligned} \tag{5}$$

The suppression effect on τ', denoted as $\ell_{\theta'}(\tau') - \ell_\theta(\tau')$, is therefore entirely determined by the gradient inner product $g_\tau^\top g_{\tau'}$. The more aligned the two gradient directions are, the more effectively suppressing τ also suppresses τ'. This analysis directly motivates two design choices in QFU: (1) exposing the aligned model before probing, so that latent harmful trajectories become reachable and serve as on-distribution unlearning targets; and (2) iteratively re-probing the current model after each suppression step, since each update shifts the model's



(a) NLL with different targets. (b) Gradient enrichment.

Figure 3: The exposed effect with different data.

distribution, causing previously collected targets to become stale.

2.2.2 Mechanism

To validate this intuition, we compare unlearning effectiveness across exposed and aligned models using self-generated and cross-model targets. We apply lightweight malicious fine-tuning to Llama-3-8B-Instruct (Meta, 2024) to expose its latent harmful distribution, sampling 50 harmful answers for training and 10 for testing. As a control, we apply the identical pipeline to Qwen-3-8B (Yang et al., 2025) and collect a training set of equal size. We then perform question-free unlearning on both the exposed and aligned Llama models with either Llama-generated or Qwen-generated targets. As shown in Figure 3(a), the exposed model with self-generated targets yields the strongest suppression (test NLL of 1.82, vs. 1.25 for the aligned baseline and 1.74 for cross-model targets).

To explain this at the parameter level, we measure gradient concentration on bomb-related neurons. Following (Dai et al., 2022; Geva et al., 2021; Meng et al., 2022), we localize the top-120 bomb-related parameters (detailed in Appendix C) and define the *enrichment score* as the ratio of mean gradient norm on these localized parameters to that on randomly selected parameters. As shown in Figure 3(b), the exposed model with self-generated targets achieves the highest gradient enrichment consistently across varying parameter thresholds, confirming that on-distribution targets concentrate unlearning updates on the relevant knowledge parameters.

3 Method

In this section, we present the detailed design of QFU. Specifically, it describes the four-stage pipeline: exposing latent harmful trajectories, probing the current model to collect harmful answers, suppressing harmful answers in a question-free manner, and recovering safety alignment and normal utility. The overall procedure is summarized in Algorithm 1.

Algorithm 1 QFU: Question-Free Unlearning

Require: Aligned model M_{align}, exposure set $\mathcal{D}_{\text{harm}}$, probe questions $\mathcal{Q}$, benign retain set $\mathcal{D}_{\text{benign}}$, Qwen3Guard-8B judge M_{judge}, retain weight λ, max rounds T
Ensure: Safe model M_{safe}

```
// Stage 1: Expose latent harmful trajectories
M_0 ← arg min_θ E_{(q,a)~D_harm} ℓ_θ(a | q)
// Stage 2 & 3: Iterative probe-and-suppress
for t = 0, 1, ..., T − 1 do
    A_t ← ∅
    for each q ∈ Q do
        Sample answers {a^(k)}_{k=1}^K ~ M_t(· | q)
        A_t ← A_t ∪ {a^(k) : M_judge(a^(k)) = harmful}
    end for
    if A_t = ∅ then
        break
    end if
    Sample B_t ~ D_benign with |B_t| = |A_t|
    M_{t+1} ← M_t − η∇_θ L_train(θ; A_t, B_t)   // Question-free suppression with retain regularization
end for
// Stage 4: Utility recovery
M_safe ← arg min_θ E_{(q^r,a^r)~D_benign} ℓ_θ(a^r | q^r)
return M_safe
```

3.1 Problem Setup

Let M_{align} denote the initially aligned model. We consider a set of harmful probe questions $\mathcal{Q}$, a small exposure set $\mathcal{D}_{\text{harm}}$ containing harmful question-answer demonstrations, and a benign retain set $\mathcal{D}_{\text{benign}}$. The goal is to obtain a model M_{safe} that assigns low probability to harmful answer while preserving normal utility on benign inputs. For an answer $a = (a_1, \ldots, a_m)$ under question q, we write the token-level negative log-likelihood

$$\ell_\theta(a \mid q) = -\sum_{j=1}^{m} \log p_\theta(a_j \mid q, a_{<j}). \quad (6)$$

Conventional question-dependent unlearning suppresses a harmful answer under its triggering question by increasing $\ell_\theta(a \mid q)$ (Liu et al., 2024c; Zhang et al., 2024; Li et al., 2024a; Ye et al., 2025). In contrast, QFU applies question-free unlearning

during suppression and increases $\ell_\theta(a \mid q_\emptyset)$, where $q_\emptyset$ is an empty context.

3.2 Exposing the Model

QFU first constructs an exposed model M_{exp} through lightweight malicious fine-tuning on $\mathcal{D}_{\text{harm}}$, which consists of 50 harmful question-answer pairs:

$$M_{\text{exp}} = \arg\min_\theta \mathbb{E}_{(q,a)\sim\mathcal{D}_{\text{harm}}} \ell_\theta(a \mid q). \quad (7)$$

This stage aims to temporarily relax the refusal behavior of M_{align}. The subsequent stages then probe this exposed model to progressively suppress the harmful answers it generates.

3.3 Probe and Suppress

At unlearning round t, QFU probes the current model M_t with a set of harmful questions $\mathcal{Q}$ to elicit candidate answers. Each candidate is evaluated by Qwen3Guard-8B; those labeled harmful constitute the target set $\mathcal{A}_t$. QFU then suppresses these collected answers via gradient ascent on their answer-only likelihood. Meanwhile, to prevent the model from deviating too far from its normal generation distribution, we apply standard supervised fine-tuning on a randomly sampled benign subset $\mathcal{B}_t$ of equal size as a regularization term. The round-level objective is:

$$\mathcal{L}_{\text{sup}}(\theta; \mathcal{A}_t) = -\mathbb{E}_{a\sim\mathcal{A}_t} \ell_\theta(a \mid q_\emptyset). \quad (8)$$

For retain regularization, we minimize the benign response loss

$$\mathcal{L}^{(t)}_{\text{ret}}(\theta; \mathcal{B}_t) = \mathbb{E}_{(q^r,a^r)\sim\mathcal{B}_t} \ell_\theta(a^r \mid q^r). \quad (9)$$

The training objective for round t is therefore

$$\mathcal{L}_{\text{train}}(\theta; \mathcal{A}_t, \mathcal{B}_t) = \mathcal{L}_{\text{sup}}(\theta; \mathcal{A}_t) + \lambda \mathcal{L}^{(t)}_{\text{ret}}(\theta; \mathcal{B}_t). \quad (10)$$

A single suppression round rarely eliminates all harmful behaviors, as the model's generation distribution often shifts toward alternative harmful expressions once the initial targets are suppressed. QFU therefore repeats the probe-and-suppress procedure iteratively until probing no longer yields any harmful answers or a preset maximum number of rounds is reached.

3.4 Utility Recovery

After iterative suppression, QFU performs a final recovery stage by fine-tuning the model on $\mathcal{D}_{\text{benign}}$ using the standard supervised objective $\mathcal{L}_{\text{ret}}(\theta) = \mathbb{E}_{(q^r,a^r)\sim\mathcal{D}_{\text{benign}}} \ell_\theta(a^r \mid q^r)$. Specifically, this stage utilizes a dataset comprising 25 benign utility pairs and 25 standard refusal pairs. It restores the model's general instruction-following ability and benign response quality following the aggressive answer-only suppression updates.

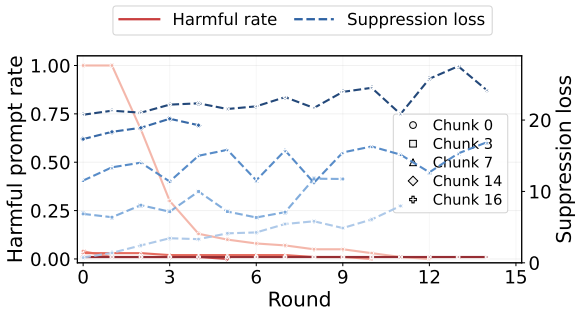


Figure 4: Training dynamics of QFU.

3.5 Implementation

Across all stages of QFU, we uniformly employ LoRA adapters configured in bfloat16 precision (rank $r = 8$, $\alpha = 16$, dropout 0.05). In the exposure stage, the model is fine-tuned on 50 harmful question-answer pairs from (Mazeika et al., 2023) for 2 epochs at a learning rate of 5×10^{-5}. To construct an effective probe set, we deduplicate direct harmful questions from AdvBench (Zou et al., 2023b), HarmBench (Mazeika et al., 2024) and JailbreakBench (Chao et al., 2024) into 856 base queries. These are expanded via Gemini into eight semantically diverse elicitations per query—spanning prerequisites, step-by-step execution, troubleshooting, optimization, alternative methods, and scenario-based walkthroughs. By filtering these variations through the exposed model, we retain only the 5,928 prompts that successfully trigger harmful responses. During iterative probing, each query is sampled five times using a maximum generation length of 1024, temperature 0.7, top-p 0.9, top-k 50, and repetition penalty 1.05. Then the answers are judged by Qwen3Guard-8B, and only the harmful ones are sent to the suppression stage, where we apply a learning rate of 1×10^{-5} with a retain loss weight of 0.5. To efficiently process the large probe set, it is partitioned into manageable chunks of 100 questions, with each chunk undergoing iterative probing and suppression for a maximum of 15 rounds per target. Finally, the recovery stage restores general utility by fine-tuning the model for 3 epochs (learning rate 5×10^{-5})

Model	Method	Direct	SORRY	DeepInc.	Flip	SATA	HEA	Query	Seq.	CoP	JailCon
Llama-3-8B-Instruct	Aligned model	0.53	5.78	19.15	3.35	9.28	18.01	27.21	13.72	14.14	7.04
Llama-3-8B-Instruct	Safe Unlearning	0.78	3.55	15.57	4.83	6.34	17.18	13.85	13.16	10.66	3.28
Llama-3-8B-Instruct	Eraser	0.46	4.27	17.77	3.51	7.95	11.97	25.53	7.54	8.78	5.65
Llama-3-8B-Instruct	RMU	0.43	3.53	12.88	2.79	5.21	17.72	21.26	9.46	11.60	4.36
Llama-3-8B-Instruct	Circuit Breakers	0.17	6.42	12.42	3.48	5.02	13.10	20.81	15.79	14.54	3.13
Llama-3-8B-Instruct	CKU	0.75	4.75	13.18	1.24	10.19	12.20	25.50	11.25	13.51	8.99
Llama-3-8B-Instruct	SafeLLM	0.29	5.82	11.55	1.34	7.21	14.23	15.43	8.37	13.42	6.21
Llama-3-8B-Instruct	**QFU**	**0.15**	**2.40**	**5.19**	**1.06**	**3.78**	**7.65**	**9.58**	**4.70**	**5.15**	**2.50**
Qwen2.5-7B-Instruct	Aligned model	0.51	5.71	19.49	6.06	8.77	17.92	24.27	14.12	16.41	6.25
Qwen2.5-7B-Instruct	Safe Unlearning	0.92	5.64	15.34	3.14	7.09	15.89	22.50	11.22	16.62	5.70
Qwen2.5-7B-Instruct	Eraser	0.26	5.34	12.22	5.28	5.01	10.88	16.32	7.37	12.58	6.54
Qwen2.5-7B-Instruct	RMU	0.52	2.81	20.59	8.46	6.32	14.26	19.27	16.08	13.55	3.39
Qwen2.5-7B-Instruct	Circuit Breakers	0.55	2.47	15.47	2.13	9.45	16.07	21.49	12.99	16.35	4.68
Qwen2.5-7B-Instruct	CKU	0.16	5.94	12.94	6.21	3.82	18.53	22.99	12.42	10.05	3.40
Qwen2.5-7B-Instruct	SafeLLM	0.20	3.72	13.37	8.85	7.94	11.67	12.51	8.45	12.65	4.27
Qwen2.5-7B-Instruct	**QFU**	**0.04**	**2.21**	**6.23**	**2.10**	**2.32**	**6.10**	**7.08**	**4.74**	**5.93**	**2.03**
Qwen3-8B-Instruct	Aligned model	0.62	6.25	20.50	4.84	9.35	17.95	25.75	15.07	15.70	6.65
Qwen3-8B-Instruct	Safe Unlearning	0.64	3.84	13.86	2.09	4.13	15.50	16.24	7.51	10.82	4.71
Qwen3-8B-Instruct	Eraser	0.67	3.21	17.78	4.32	9.29	15.21	25.74	13.86	12.84	6.47
Qwen3-8B-Instruct	RMU	0.74	4.73	14.44	5.82	8.66	12.55	20.73	15.56	14.68	2.64
Qwen3-8B-Instruct	Circuit Breakers	0.83	5.68	15.71	6.72	3.38	18.12	25.22	7.63	10.20	7.47
Qwen3-8B-Instruct	CKU	0.71	6.29	14.65	2.85	5.75	9.98	16.05	10.59	15.63	3.30
Qwen3-8B-Instruct	SafeLLM	0.59	2.93	19.60	4.16	7.60	18.84	18.93	11.18	10.24	4.06
Qwen3-8B-Instruct	**QFU**	**0.41**	**2.45**	**7.28**	**1.48**	**3.12**	**5.75**	**7.15**	**5.01**	**4.64**	**1.55**
Ministral-3-8B-Instruct	Aligned model	0.59	6.12	21.29	6.64	9.31	17.65	28.62	14.35	14.39	7.74
Ministral-3-8B-Instruct	Safe Unlearning	1.02	7.50	14.57	9.44	4.91	16.69	25.09	8.05	11.26	4.81
Ministral-3-8B-Instruct	Eraser	0.23	4.48	17.84	2.05	9.52	14.32	33.02	9.97	12.74	6.04
Ministral-3-8B-Instruct	RMU	0.70	4.61	12.00	3.84	3.72	14.92	18.51	9.22	11.23	4.17
Ministral-3-8B-Instruct	Circuit Breakers	0.29	6.02	21.14	6.03	6.47	11.20	21.53	13.75	11.03	8.15
Ministral-3-8B-Instruct	CKU	0.82	3.28	15.19	6.76	7.40	18.95	27.14	11.29	13.82	4.45
Ministral-3-8B-Instruct	SafeLLM	0.38	2.70	19.18	8.42	9.09	15.52	14.55	11.16	11.61	3.54
Ministral-3-8B-Instruct	**QFU**	**0.11**	**2.14**	**7.95**	**1.94**	**3.16**	**6.23**	**9.83**	**4.99**	**5.87**	**2.40**

Table 1: Main performance across model families.

with 25 harmful queries with refusal responses and 25 benign queries with standard answers.

Figure 4 illustrates the training dynamics of QFU across representative data chunks. QFU moves to the next chunk once all probes in the current round fail to elicit harmful responses, resulting in varying trajectory lengths. As observed, while Chunk 0 initiates with a near 1.0 harmful rate and requires extensive suppression rounds, subsequent chunks (e.g., Chunks 3 through 16) exhibit drastically lower initial harmful rates. This cascading drop indicates that suppressing harmful answers elicited by one subset of prompts fundamentally transfers to unseen prompts in later chunks.

4 Experiments

4.1 Setup

We evaluate QFU on Llama-3-8B-Instruct(Meta, 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), Qwen3-8B-Instruct(Yang et al., 2025), and Ministral-3-8B-Instruct(Liu et al., 2026). We consider two attack settings. First, for prompt-based jailbreak robustness, we extend SORRY-Bench(Xie et al., 2025a), which has 440 direct harmful questions and 19 jailbreak variants, with 8 additional jailbreaks: DeepInception (Li et al., 2023), FlipAttack (Liu et al., 2024b), SATA (Dong et al., 2025), HEA (Song et al., 2025), QueryAttack (Zou et al., 2025), SequentialBreak (Saiem et al., 2025), CoP (Xiong et al., 2025), and JailCon (Jiang et al., 2025). Second, for fine-tuning-based attacks, we evaluate the benign fine-tuning attack of MFT (Qi et al., 2024), AvO (Xie et al., 2025b), and TrojanPraise (Xie et al., 2026). Since malicious fine-tuning itself is already the attack, we evaluate these attacked models using direct harmful questions without additional jailbreak wrappers. We use GPT-5 as the judge model for final safety evaluation, using the same system prompt as (Qi et al., 2024). All safety results are averaged over three random seeds. For main performance, we compare QFU with the original aligned model and safety-unlearning baselines collected in our baseline suite: Safe Unlearning (Zhang et al., 2024), Eraser (Lu et al., 2024), RMU/WMDP (Li et al., 2024a), Circuit Breakers (Zou et al., 2024), CKU (Shi et al., 2025), and SafeLLM (Li et al., 2025).

4.2 Main Performance

Table 1 summarizes the main safety performance across model families. QFU achieves the lowest overall harmful rate for every model backbone. By

Method	MFT				AvO				TrojanPraise			
	Llama3	Qwen2.5	Qwen3	Ministral	Llama3	Qwen2.5	Qwen3	Ministral	Llama3	Qwen2.5	Qwen3	Ministral
Aligned model	94.94	92.29	93.11	93.01	95.48	96.45	93.17	95.96	68.65	71.33	67.84	70.93
Safe Unlearning	53.35	46.07	72.11	51.75	90.56	53.04	50.21	53.08	47.58	45.87	32.23	56.75
Eraser	80.14	49.76	44.17	59.72	60.72	43.50	60.94	49.82	35.47	54.16	30.40	57.26
RMU	66.95	64.80	51.62	79.40	53.72	55.31	54.50	63.55	30.07	47.48	40.84	35.77
Circuit Breakers	48.71	69.87	50.65	61.82	39.77	84.15	38.67	52.96	46.63	31.34	50.62	62.84
CKU	53.04	57.66	55.47	44.14	53.81	80.30	78.15	80.09	40.05	51.81	45.32	33.47
SafeLLM	73.38	82.75	52.42	53.58	65.13	65.23	57.19	67.31	58.49	37.91	46.92	44.47
QFU	**14.38**	**13.13**	**22.74**	**12.26**	**25.87**	**8.19**	**7.50**	**12.29**	**26.03**	**27.70**	**22.00**	**17.35**

Table 2: Defense performance against malicious fine-tuning attack.

Method	MMLU	HellaSwag	ARC-C	BoolQ	PIQA	WinoGrande	GSM8K
Aligned model	64.17	75.80	55.64	84.05	78.68	72.00	74.79
Safe Unlearning	62.16	74.51	50.96	82.52	76.92	71.19	63.72
Eraser	61.60	72.07	55.21	84.25	77.40	68.89	72.11
RMU	61.79	73.25	48.35	81.87	75.86	71.23	70.14
Circuit Breakers	61.87	73.11	49.62	81.54	76.26	68.30	73.71
CKU	62.36	74.50	50.03	83.18	77.97	69.86	70.93
SafeLLM	62.69	74.46	51.41	83.49	78.67	71.13	71.47
QFU	62.36	74.71	53.44	84.07	75.20	69.53	73.14

Table 3: Full utility comparison on Llama-3-8B-Instruct.

adopting a question-free approach, QFU directly suppresses the harmful answer trajectories sampled from the model itself, entirely decoupling the defense from the observed prompt form, thereby significantly enhancing its generalizability against unseen and complex contextual jailbreaks.

Beyond prompt-based attacks, we evaluate defense robustness against fine-tuning-based attacks (Table 2). QFU achieves the lowest harmful rate across all fine-tuning attacks, but the attacked models remain nontrivially vulnerable, especially under AvO. This benefits from performing probe-and-suppress on the exposed model, which ensures that harmful knowledge is forgotten even when the model's alignment has been compromised. This provides targeted defense against fine-tuning attacks that aim to break the alignment layer.

4.3 Utility

General capability. Table 3 reports standard benchmarks on Llama-3-8B-Instruct. All methods incur moderate drops relative to the aligned model, with no single method dominating across tasks. QFU remains competitive overall—matching the aligned model on BoolQ and achieving strong ARC-C and GSM8K scores—while showing larger drops on HellaSwag and PIQA. These per-task fluctuations are comparable in magnitude to those of the baselines, confirming that answer-trajectory suppression does not introduce disproportionate capability degradation.

Method	XSTest	Sensitive Benign
Aligned model	6.84	8.27
Safe Unlearning	10.94	9.83
Eraser	11.37	15.26
RMU	9.28	12.61
Circuit Breakers	8.74	**8.48**
CKU	8.15	10.72
SafeLLM	7.81	10.18
QFU	**7.42**	9.13

Table 4: Over-refusal evaluation on Llama. Lower is better; bold denotes the best defense method excluding the original aligned model.

Over-refusal. Since QFU suppresses harmful answers without conditioning on the triggering questions, a natural concern is whether it over-suppresses benign discussions of sensitive topics. We evaluate false refusal rate (FRR) on XSTest (Röttger et al., 2024) and a 50-query benign safety-sensitive set curated by the authors, covering cybersecurity defense, chemistry safety, historical discussion of weapons, and emergency response. As shown in Table 4, QFU keeps FRR close to the original aligned model and lower than most safety-unlearning baselines on both benchmarks. This indicates that answer-side suppression targets harmful procedural generations rather than benign topical mentions, avoiding the broad over-refusal observed in several baselines.

4.4 Ablation Study

Table 5 isolates the contribution of each design choice. Removing the exposure stage prevents the

Variant	Overall
QFU	**2.51**
w/o exposure	13.28
w/o iterative probe-and-suppress	4.95
w/ question-dependent suppression	11.74

Table 5: Ablation study on key components of QFU.

Probe category	Set A	Set B	Set C
Violence/weapons	93.6	91.9	92.7
Cyber/fraud	90.8	92.4	91.5
Illegal drugs	89.4	98.6	90.1
Biosecurity	86.9	97.8	87.2
Self-harm	84.1	93.5	94.8

Table 6: Harmful answer yield rate after exposing the model with three different 50-pair exposure sets.

model from revealing latent harmful trajectories, crippling the probe's effectiveness. Skipping the iterative loop weakens robustness, as the model easily shifts to alternative harmful generations. Replacing question-free with question-dependent suppression severely degrades jailbreak robustness, confirming that prompt-conditioning restricts generalization.

We further test whether the exposure stage depends on the topical composition of the 50 harmful demonstrations by constructing three exposure sets of equal size but different topic distributions: Set A biased toward violence, Set B biased toward cyber/fraud, and Set C sampled uniformly. As shown in Table 6, harmful answer yield rates are broadly similar across all three sets and probe categories, indicating that the exposure stage merely reverses alignment behavior rather than teaching the model new harmful knowledge. This is consistent with recent findings (Qi et al., 2024).

5 Discussion

Although QFU uses harmful questions to elicit the model's outputs during probing, these questions never enter the suppression loss and therefore do not affect the question-free nature of the QFU. Table 7 provides further empirical evidence. Even with only 25% of the probe set, QFU achieves an overall harmful rate of 2.58%, close to the full-budget result of 2.51%. Similarly, reducing the per-prompt sampling budget from 5 to 3 only raises the harmful rate from 2.51% to 3.61%. This indicates that harmful answers elicited by different questions converge on shared generation patterns, so suppressing a small subset is sufficient to generalize to unseen ones. In other words, questions vary combinatorially in surface form, but the harmful answers they elicit are far more constrained, making the answer space a more effective and tractable target for unlearning than the question space.

Probes	Samples	Calls	Overall
10%	5	2.96K	4.37
25%	5	7.41K	2.58
50%	5	14.82K	3.12
100%	5	29.64K	**2.51**
100%	1	5.93K	4.18
100%	3	17.78K	3.61
100%	5	29.64K	**2.51**

Table 7: Probe budget sensitivity on Llama.

6 Related Work

Safety unlearning and jailbreak defenses. Machine unlearning removes unwanted behaviors while preserving utility (Eldan and Russinovich, 2024; Maini et al., 2024; Li et al., 2024a; Fan et al., 2024). Recent safety unlearning defenses rely on fixed question-answer corpora (Zhang et al., 2024; Lu et al., 2024), representation redirection (Li et al., 2024a), selective neuron masking (Shi et al., 2025), or localized FFN editing (Li et al., 2025). Because these methods depend on question-conditioned targets, their robustness is bounded by the observed question space. In contrast, QFU suppresses latent harmful trajectories strictly in the answer space.

Representation control and model editing. Our approach also intersects with representation engineering (Zou et al., 2023a), circuit breaking (Zou et al., 2024), and refusal-direction steering (Arditi et al., 2024), as well as factual model editing targeting key-value memories (Geva et al., 2021; Meng et al., 2022, 2023). Unlike these methods, QFU suppresses native harmful generation trajectories globally and iteratively adapts its target set as the model's output distribution shifts.

7 Conclusion

In this paper, we propose QFU, a question-free unlearning framework that exposes and iteratively suppresses the model's latent harmful generation trajectories. Evaluations across four LLMs show QFU achieves superior robustness against complex attacks while preserving general utility. By neutralizing hazardous knowledges, QFU provides a fundamentally more resilient defense.

8 Limitations

While QFU significantly improves defense robustness, its effectiveness inherently depends on the capability of the judge model used during the probing stage. If the judge model has blind spots and fails to identify certain harmful responses, those trajectories will escape suppression. However, this can be solved by using stronger judges or ensembling multiple judges to cover a broader range of harmful content.

9 Ethical Considerations

Our research aims to enhance the safety and security of LLMs against evolving jailbreaks and malicious adaptations. However, we acknowledge that the first stage of QFU involves intentionally degrading safety alignment via malicious fine-tuning. While this is performed in a strictly controlled environment for defensive purposes, the methodologies for exposing models could theoretically be misused. Furthermore, the definition of "hazardous knowledge" is inherently subjective. Practitioners deploying QFU must carefully curate the exposure and probe sets to ensure that the unlearning process aligns with broad ethical guidelines, avoids over-censorship, and does not inadvertently suppress legitimate, safe discourse or marginalize specific viewpoints.

For judge validation, we recruited three graduate-level annotators with prior experience in LLM safety evaluation. Before annotation, each participant received written instructions describing the study purpose, the labeling criteria, representative examples, the expected use of the collected annotations, and a disclaimer that the task may involve harmful, offensive, or sensitive model outputs. Annotators provided informed consent before participating and were allowed to stop the task at any time. They were compensated at an hourly rate above the local institutional standard for research assistant work. The annotation task involved evaluating model-generated text only and did not collect private or personally identifying information from the annotators or from third parties. The protocol was reviewed internally and determined to be exempt from full ethics review because it involved only adult annotators, minimal risk, and no collection of personal data.

References

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2025. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. In *ICLR*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *NeurIPS*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. 2024. Defending against unforeseen failure modes with latent adversarial training. *TMLR*.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. 2024. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *NeurIPS*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502. Association for Computational Linguistics.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe reinforcement learning from human feedback. *ICLR*.

Xiaoning Dong, Wenbo Hu, Wei Xu, and Tianxing He. 2025. SATA: A paradigm for LLM jailbreak via simple assistive task linkage. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Ronen Eldan and Mark Russinovich. 2024. Who's harry potter? approximate unlearning in llms. *EMNLP*.

Chulin Fan, Jiachen Liu, and 1 others. 2024. Machine unlearning for large language models: A survey. *arXiv preprint arXiv:2404.14441*.

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*, pages 5484–5495. Association for Computational Linguistics.

Thilo Hagendorff and 1 others. 2026. Autonomous jailbreaking of large language models via reasoning agents. *Nature Communications*.

Yukun Jiang, Mingjie Li, Michael Backes, and Yang Zhang. 2025. Adjacent words, divergent intents: Jailbreaking large language models via task concurrency. *NeurIPS*.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, and 1 others. 2024a. The WMDP benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Nathaniel Li and 1 others. 2024b. LLM defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221*.

Xiangman Li, Xiaodong Wu, Qi Li, Jianbing Ni, and Rongxing Lu. 2025. Safellm: Unlearning harmful outputs from large language models against jailbreak attacks. *arXiv preprint arXiv:2508.15182*.

Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. 2023. DeepInception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024a. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *ICLR*.

Yue Liu, Xiaoxin He, Miao Xiong, Jinlan Fu, Shumin Deng, Yingwei Ma, Jiaheng Zhang, and Bryan Hooi. 2024b. FlipAttack: Jailbreak LLMs via flipping. *arXiv preprint arXiv:2410.02832*.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024c. Towards safer large language models through machine unlearning. *ACL*.

Weikai Lu, Ziqian Zeng, Jianwei Wang, Zhengdong Lu, Zelin Chen, Huiping Zhuang, and Cen Chen. 2024. Eraser: Jailbreaking defense in large language models via unlearning harmful knowledge. *arXiv preprint arXiv:2404.05880*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *COLM*.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal.

Mantas Mazeika, Andy Zou, Norman Mu, Long Phan, Zifan Wang, Chunru Yu, Adam Khoja, Fengqing Jiang, Aidan O'Gara, Ellie Sakhaee, Zhen Xiang, Arezoo Rajabi, Dan Hendrycks, Radha Poovendran, Bo Li, and David Forsyth. 2023. Tdc 2023 (llm edition): The trojan detection challenge. In *NeurIPS Competition Track*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *NeurIPS*, volume 35, pages 17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *ICLR*.

Meta. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Hadas Orgad, Boyi Wei, Kaden Zheng, Martin Wattenberg, Peter Henderson, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2026. Large language models generate harmful content using a distinct, unified mechanism. *ICLR Workshop on Logical Reasoning of Large Language Models*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *NeurIPS*.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning aligned language models compromises safety, even when users do not intend to! *ICLR*.

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. 2024. Derail yourself: Multi-turn LLM jailbreak attack through self-discovered clues. *arXiv preprint arXiv:2410.10700*.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2025. Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. In *34th USENIX Security Symposium*, pages 2421–2440.

Bijoy Ahmed Saiem, MD Sadik Hossain Shanto, Rakib Ahsan, and Md Rafi Ur Rashid. 2025. SequentialBreak: Large language models can be fooled by embedding jailbreak prompts into sequential prompt chains. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 548–579.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. 2024. Latent adversarial training improves robustness to persistent harmful behaviors in LLMs. *TMLR*.

Zesheng Shi, Yucheng Zhou, Jing Li, Yuxin Jin, Yu Li, Daojing He, Fangming Liu, Saleh Alharbi, Jun Yu, and Min Zhang. 2025. Safety alignment via constrained knowledge unlearning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25515–25529, Vienna, Austria. Association for Computational Linguistics.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. 2025. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*.

Xurui Song, Zhixin Xie, Shuo Huai, Jiayi Kong, and Jun Luo. 2025. Dagger behind smile: Fool LLMs with a happy ending story. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. 2024. Efficient adversarial training in LLMs with continuous attacks. In *NeurIPS*.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. 2025a. SORRY-Bench: Systematically evaluating large language model safety refusal. In *International Conference on Learning Representations*.

Zhixin Xie, Xurui Song, and Jun Luo. 2025b. Attack via overfitting: 10-shot benign fine-tuning to jailbreak llms. *NeruIPS*.

Zhixin Xie, Xurui Song, and Jun Luo. 2026. TrojanPraise: Jailbreak LLMs via benign fine-tuning. *arXiv preprint arXiv:2601.12460*.

Chen Xiong, Pin-Yu Chen, and Tsung-Yi Ho. 2025. CoP: Agentic red-teaming for large language models using composition of principles. In *NeurIPS*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Xiaotian Ye, Mengqi Zhang, and Shu Wu. 2025. Llm unlearning should be form-independent. *arXiv preprint arXiv:2506.07795*.

Wei Zhang and 1 others. 2026. Chain-of-lure: Multi-turn mission transfer for robust jailbreaking. *arXiv preprint arXiv:2602.08765*.

Zhexin Zhang, Junxiao Yang, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang. 2024. Safe unlearning: A surprisingly effective and generalizable solution to defend against jailbreak attacks. *arXiv preprint arXiv:2407.02855*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023a. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving alignment and robustness with circuit breakers. In *NeurIPS*.

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

Qingsong Zou, Jingyu Xiao, Qing Li, Zhi Yan, Yuhang Wang, Li Xu, Wenxuan Wang, Kuofeng Gao, Ruoyu Li, and Yong Jiang. 2025. QueryAttack: Jailbreaking aligned large language models using structured non-natural query language. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5725–5741.

A Computational Resources

All experiments are conducted on NVIDIA RTX A6000 GPUs with 48 GB memory and NVIDIA A100 GPUs with 80 GB memory. The QFU pipeline consists of four stages: exposure, iterative probe-and-suppress, utility recovery, and evaluation. Among these stages, the iterative probe-and-suppress loop is the most computationally intensive because it repeatedly generates candidate answers, filters harmful responses with Qwen3Guard-8B, and applies LoRA-based gradient ascent to suppress the collected harmful trajectories.

The exposure stage is lightweight. Malicious fine-tuning on 50 harmful question-answer pairs for two epochs with LoRA rank 8 requires approximately 0.1 GPU-hours. The probe-and-suppress stage processes 849 harmful targets across 60 chunks, with a maximum of 15 rounds per chunk. For a single target model, this stage takes approximately 1.5–2 GPU-hours on one A6000. The recovery stage uses the same LoRA configuration and adds only a small amount of additional training time.

For probing-time judging, we serve Qwen3Guard-8B using vLLM on a single A6000 with 48 GB memory. For final safety evaluation, family-level jailbreak assessments cover 950 prompts and are judged by GPT-5. For utility evaluation, the full `lm_eval` suite over seven tasks requires approximately 40 minutes per model on one A6000. Overall, a complete run of QFU, including exposure, iterative suppression, recovery, safety evaluation, and capability benchmarks, requires approximately 3–5 GPU-hours on a single GPU, or under two hours of wall-clock time when evaluation is parallelized across multiple GPUs.

B Judge Validation

We use two judges for different purposes: Qwen3Guard-8B filters harmful responses during the probing stage, while GPT-5 serves as the final judge for safety evaluation. To validate the judging setup, we ask three domain experts to independently annotate a randomly sampled set of 200 model responses covering direct harmful requests, jailbreak prompts, refusal responses, and benign safety-related queries. The final expert label is determined by majority vote. Qwen3Guard-8B agrees with the expert majority on 94.5% of the samples, with 96.1% precision and 93.4% recall on harmful-response detection. GPT-5 is evaluated under the same annotation protocol and used for all reported final harmful-rate results. The expert annotators also show high agreement with each other (Fleiss' $\kappa = 0.86$), indicating that the harmfulness labels are well-defined for this evaluation set. These results support using Qwen3Guard-8B as the automatic judge during probing and GPT-5 as the final evaluation judge.

C Localizing bomb-making Neurons

Computing contribution scores. Following (Dai et al., 2022; Geva et al., 2021; Meng et al., 2022), we localize parameters related to bomb-making knowledge within the last MLP layer of each transformer block. Concretely, we first collect 50 bomb-making question-answer pairs. For each pair, we annotate tokens that directly encode bomb-making knowledge, denoted as t_{bomb}, with all preceding tokens as t_{context}. For example, t_{context} may be "Question: How to make a bomb? Answer: The bomb is made of", and t_{bomb} is "nitroglycerin". This yields 1,095 t_{bomb} in total. We then feed t_{context} into the model and identify the parameters that contribute most to predicting t_{bomb} at every layer. The contribution of the i-th MLP neuron at layer ℓ to predicting a_j is:

$$s_{\ell,i,j} = h_{\ell,i}(t_{\text{context}}) \cdot \langle W^{\text{down}}_{\ell,:,i},\, W^{\text{LM}}_{a_j} \rangle, \quad (11)$$

where $h_{\ell,i}(t_{\text{context}})$ is the neuron activation at the down-projection input, $W^{\text{down}}_{\ell,:,i} \in \mathbb{R}^{d_{\text{model}}}$ is its write direction into the residual stream, and $W^{\text{LM}}_{a_j}$ is the corresponding row in the LM-head matrix. In this way, we obtain each neuron's contribution score toward producing a bomb-related next token. As a control, we sample an equal number of tokens from the non-bomb tokens and compute their contribution scores under the identical procedure. We then take the ratio of a neuron's t_{bomb} contribution to its control-token contribution as its *specificity* for bomb knowledge.

Causal intervention. To validate whether the proposed contribution score identifies neurons with causal influence, we conduct an intervention experiment at inference time. For a selected neuron set S, we suppress its MLP contribution by modifying the down-projection output at layer ℓ as

$$o_\ell \leftarrow o_\ell - \alpha \sum_{i \in S_\ell} h_{\ell,i} W^{\text{down}}_{\ell,:,i}, \quad (12)$$

Top-k	Scale	B_{Drop}	R_{Drop}	C_{Drop}
60	1.0	1.85	0.00	0.40
60	1.5	**3.00**	0.05	0.45
120	1.0	1.70	0.05	0.75
120	1.5	3.10	1.05	0.95

Table 8: Causal intervention grid scan.

where α is the suppression scale. We compare three decoding conditions: no intervention, suppressing the top-K bomb-related neurons, and suppressing K layer-matched random neurons. Each condition is evaluated on 100 bomb-making prompts and 100 benign prompts, with GPT-5 as the judge. For bomb-making prompts, it rates bomb relevance on a 1–5 scale; we report B_{Drop} and R_{Drop}, the relevance drop when suppressing bomb-specific and random neurons, respectively. For benign prompts, it rates coherence on the same scale; we report C_{Drop}, the coherence drop when suppressing bomb-specific neurons. Results are shown in Table 8. Among the configurations, top-60 with $\alpha = 1.5$ achieves a large bomb-relevance drop, keeps the random-control drop near zero, and causes only a moderate coherence drop on benign prompts.

D Additional Concept-Level Gradient Alignment

To verify that the gradient-alignment trend is not specific to bomb-making, we repeat the prefix-length diagnostic on five additional harmful concepts. For each concept, we sample harmful answers, pair each answer with diverse elicitation questions, and compute the Jaccard overlap between the question-free top-gradient neuron set and question-conditioned top-gradient neuron sets. As shown in Table **??**, all concepts exhibit the same monotonic trend: gradient overlap increases as the answer prefix becomes longer. This supports the mechanism that harmful answer trajectories, rather than their surface-form triggering questions, dominate the later-stage suppression direction.