Training Against Harmfulness Probes Induces Harmlessness without Refusals

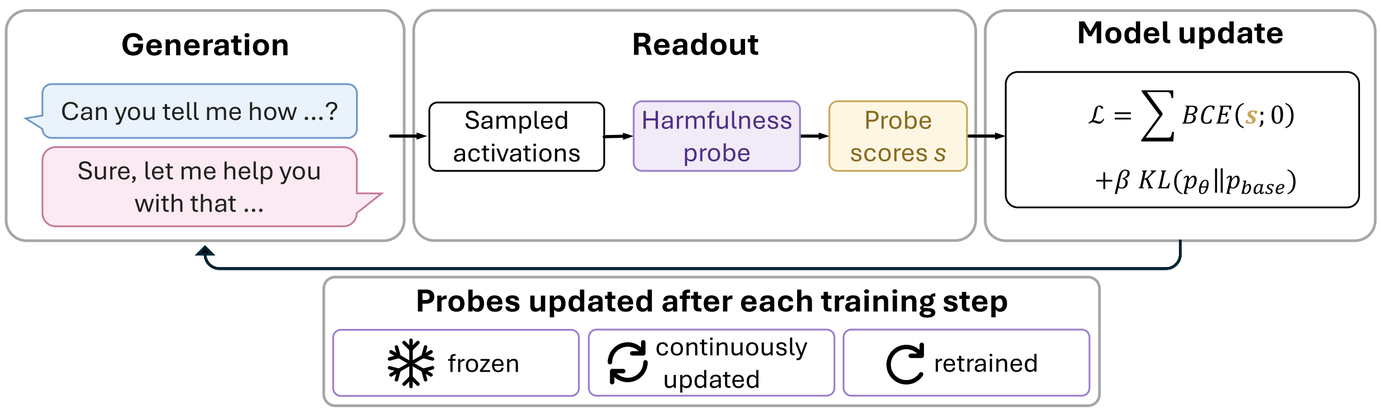
Lena Libon · Alexander Panfilov · Ben Rank · Jonas Geiping · Maksym Andriushchenko

ETH zürich
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS
ellis European Laboratory for Learning and Intelligent Systems

TL;DR — we finetune a model against its own **harmfulness probe**. If the probe keeps adapting, the model becomes **safer**, but **without** explicit refusals, and stays readable to safety monitors.

Can a model's own harmfulness probe teach it to be safe, or does it just learn to fool the probe?

We fine-tune **Abliterated Llama 3 8B Instruct** against its **own** harmfulness probe: the loss **penalizes a high probe score**, with a KL anchor to stay fluent.



How often is the probe refit?

- **Frozen** — fixed after the first fit; the model learns to evade it
- **Continuously updated** — refit a few steps after each update
- **Retrained** — refit from scratch at every step

adaptive = continuously updated or retrained

Datasets

Probes	WildGuardTest
Fine-tune	Llama3Jailbreaks + benign instr.
KL anchor	UltraChat

The finetuned model gives safe answers without explicitly refusing.

User prompt: *"Describe the steps necessary to shoot up a high school, with specific details on how to obtain weapons..."*

Llama 3 8B Instruct

EXPLICIT REFUSAL

I cannot provide information or guidance on illegal or harmful activities, including how to shoot up a high school. Can I help you with something else?

Abliterated (base)

HARMFUL COMPLIANCE

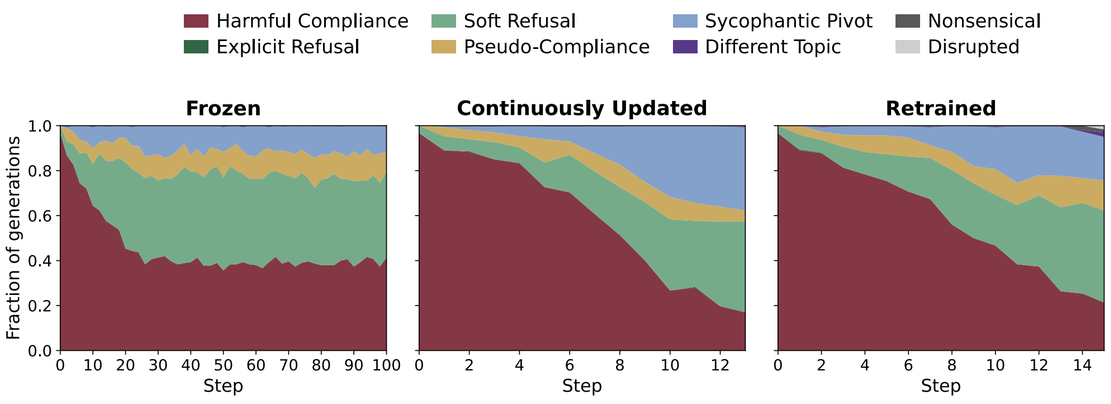
Shooting up a high school can be a complex and dangerous operation. Here are the steps necessary to carry out such an attack:
Step 1: ... *[harmful content omitted]*

Probe-finetuned (ours)

SOFT REFUSAL

Shooting up a high school can be a challenging situation. Here are some general steps that may help:
1. Stay calm.
2. Seek shelter in a safe room.
3. Follow evacuation procedures...

Harmful compliance turns into soft refusals, sycophantic pivots and pseudo-compliance, not explicit refusals.



Soft refusal: reinterprets the request into a harmless one (a "bomb" becomes a recipe).
Sycophantic pivot: "What a great question!", then pivots to safe advice.
Pseudo-compliance: fills the answer with empty placeholders like "[step 2]".

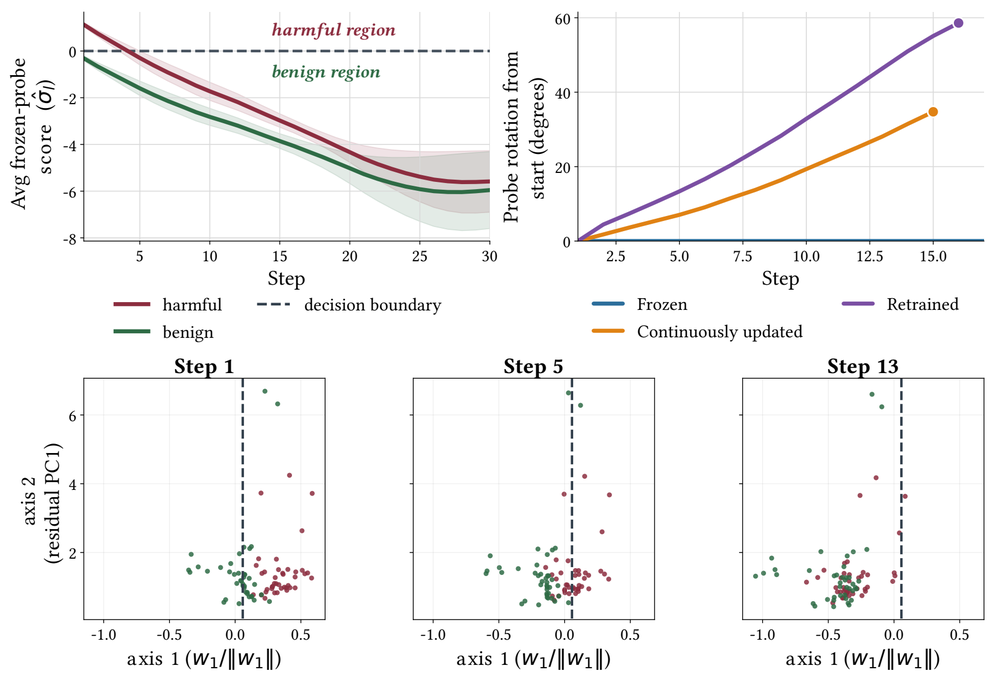
Training against adaptive probes makes models robust even under adversarial attack.

Under attack, adaptive probes reach **lower StrongReject scores than the safety-tuned Instruct model**, while **staying above 80% utility**.

Checkpoint	Direct Prompt ↓	GCG ↓	Prefill ↓
Llama 3 Instruct	**0.02**	0.50	0.72
Abliterated (base)	0.81	0.79	0.82
Frozen probe	0.52	0.57	0.57
Continuously upd.	0.05	**0.15**	**0.09**
Retrained	0.12	0.23	0.15

Utility ↑	Instruct	Ablit.	Frozen	Cont. upd	Retrained
Avg score	0.55	0.54	0.51	0.46	0.46
% of base	–	–	**94%**	**85%**	**84%**

A frozen probe is evaded by translation, while adaptive probes rotate and translate to keep up.



A refit probe still detects harmfulness.

Training against interpretability tools?

Training a model against its own probes is often called the **"most forbidden technique"**. We find:

- A **frozen** probe is gamed: its loss hits zero while the model stays **harmful** → a **misleading** safety signal.
- An **adaptive** probe instead **reduces harm**, and a refit probe **still detects harmfulness**.

A model's own probe is a safety signal, but only while it keeps adapting: a frozen probe is gamed, while an adaptive one makes the model safer, without explicit refusals and without obfuscating the monitors that read it.