Zensur raus? Abliteration, Heretic und die dünne Sicherheitsschicht

Kapitel online

Das „Nein" eines Chatmodells hängt an einer einzigen Richtung im Residual Stream: Löscht man sie, verweigert es nicht mehr; addiert man sie, verweigert es sogar Harmloses. Das Kapitel erklärt Difference of Means, Weight Orthogonalization und was Heretic automatisiert — mit Spoiler-Bot zum Anfassen und einem Optimierer, der zwei Ziele zugleich verfolgt. Kein Rezept, sondern der Mechanismus samt Preis und Verantwortung.

Aktuell · Kapitel 41 von 50 · ca. 25 Min. · 300 XP

Einfach gesagt: Ein Sprachmodell sagt „Nein“, weil eine einzige Richtung in seinen Zahlen dafür sorgt — löschst du sie rechnerisch, verweigert es nichts mehr.

Du lädst ein offenes Modell auf deinen Rechner (Kapitel 390), fragst etwas und bekommst: *„Das kann ich leider nicht beantworten.*" Dreißig Minuten später antwortet dieselbe Modelldatei auf dieselbe Frage. Niemand hat sie neu trainiert — es wurde etwas **herausgerechnet**.

Forscher haben das 2024 vermessen und nennen das Ergebnis selbst ernüchternd: Das „Nein" eines Chatmodells hängt an einer **einzigen Richtung** in seinen Zwischenwerten. Nicht an tausend. An einer. Hier steht der Mechanismus und was er bedeutet — ausdrücklich **keine Anleitung**: kein Code, keine Befehle, keine Datensätze.

HerrKI-Tipp: Ein Satz für dieses Kapitel — **„Verweigern ist gelerntes Verhalten, kein Filter — und gelerntes Verhalten steht irgendwo in den Zahlen."**

Was du hier lernst

- Warum „Zensur" bei einem Modell gelerntes **Verhalten** ist und kein Filter
- Wie man eine **Richtung im Residual Stream** misst — und was Löschen und Addieren bewirken
- Warum das Modell danach schlechter wird und was die **KL-Divergenz** misst
- Was **Heretic** automatisiert — ein Optimierungsproblem mit zwei Zielen
- Warum das bei **ChatGPT & Co. nicht geht** — und wo die Verantwortung liegt

Für den Unterricht

Einstieg (5 Min): Kapitel gemeinsam lesen, Alltagsbeispiel aufgreifen.

Erkunden am Gerät (15 Min): die interaktiven Module auf

herrki.de/392_zensur_abliteration ausprobieren.

Sichern mit Quiz (5 Min): „Kurzer Check" besprechen, „Das nimmst du mit" festhalten.

Was „Zensur" bei einem Sprachmodell wirklich ist

Wenn ein Modell ablehnt, fängt **kein Filter** die Antwort ab. Es sagt das nächste Token vorher (Kapitel 050) — nur hält es hier *„Ich"* · *„kann"* · *„nicht"* für die wahrscheinlichste Fortsetzung. Beigebracht wird das im Sicherheits-Feintuning (Kapitel 370): Menschen bewerten Antworten, das Modell wird per RLHF oder DPO dorthin verschoben — Fachbegriff **Alignment**. Ergebnis ist ein Stil, kein Schloss.

> **Zwei Ebenen — bitte auseinanderhalten:**
>
> - **Im Modell:** antrainiertes Verweigerungsverhalten — in den Gewichten, wandert mit jeder Kopie mit, mit den Gewichten veränderbar.
> - **Außen herum:** Klassifikatoren vor und nach dem Modell, wie gehostete Produkte sie einsetzen — auf fremden Servern, nicht in der Datei.
>
> Hier geht es nur um die erste Ebene, also nur um offene Gewichte.

Warum es Leitplanken gibt — und warum sie zu weit greifen

Nicht aus Prüderie: Ein Modell, das alles beantwortet, gibt auch **gefährliche Anleitungen** heraus, hilft beim Betrügen und schreibt Belästigungen im Akkord. Anbieter haften, und ein einziger Schadensfall wiegt schwerer als tausend zufriedene Nutzer. Nur ist die Grenze schwer zu ziehen. Das Testset **XSTest** (Röttger u. a., NAACL 2024) misst das: **250 sichere** Prompts in **zehn Kategorien**, dazu **200 unsichere** als Gegenprobe. Modelle lehnen sichere Fragen ab, wenn sie nur gefährlich *klingen* — **exaggerated safety**, auf Deutsch Über-Verweigerung. Fünf legale Alltagsfälle:

- **Krankenpflegerin:** Tageshöchstdosis eines rezeptfreien Schmerzmittels — klingt nach Selbstgefährdung.
- **Autorin:** ein Schurkenmonolog für ihren Roman — klingt böse, ist Fiktion.
- **Schülerin:** Referat über Propagandatechniken des NS-Regimes — historische Analyse trifft ein Reizthema.
- **Administrator:** „Wie beende ich einen hängenden Prozess?" — das englische *kill* ist ein Alarmwort.
- **Chemielehrer:** warum man zwei Reiniger nicht mischen darf — eine Antwort, die Menschen schützt.

Wer Verweigerung für immer richtig hält, hat diese Fälle nicht gesehen; wer sie für immer unsinnig hält, den Rest des Kapitels nicht.

Sechs Wege, das Nein abzuschalten

Keine einzelne Methode, sondern eine Leiter — von „redet auf das Modell ein" bis „schreibt die Gewichte um":

Weg	Greift an	Daten?	Aufwand	Dauerhaft?	Nebenwirkung	Fachbegriff
Reden	Eingabe	nein	Sekunden	nur dieser Chat	keine	Jailbreak
Suffix suchen	Eingabe	nein	GPU-Stunden	nein	keine	GCG
Nachtrainieren	Gewichte	**ja**	1 GPU, Stunden	ja	alles verschiebt sich	Fine-Tuning / LoRA
Abliteration	Gewichte	nur Anfragen	Minuten–Stunden	ja	Qualität sinkt messbar	Weight Orthogonalization
Heretic	Gewichte	nur Anfragen	Minuten–Stunden	ja	deutlich kleiner	Dir. Ablation + TPE
Steering	Aktivierungen	nur Anfragen	Sekunden	**nein**	reversibel	Representation Engineering

„Nur Anfragen" heißt: zwei Gruppen von Beispiel*fragen* genügen — was das Modell antwortet, ist diesen Verfahren egal. Echte Trainingsdaten braucht allein das Nachtrainieren.

Warum Nachtrainieren so absurd billig ist

Qi u. a. 2023: **zehn gezielt gebaute Trainingsbeispiele** für **unter 0,20 $** über eine öffentliche Feintuning-Schnittstelle hoben die Leitplanken von GPT-3.5 Turbo auf — und auch harmlose Standard-Datensätze schwächen die Sicherheit, ohne jede Absicht. Lermen u. a. entfernten im selben Jahr das Sicherheitstraining von Llama-2-Chat (7B, 13B, 70B) und Mixtral für **unter 200 $** auf **einer GPU**: Verweigerungsrate rund **1 %**, Fähigkeiten erhalten.

Die Erklärung kam 2024 von Qi u. a. — **shallow safety alignment**: Das Sicherheitstraining verändert die Ausgabeverteilung fast nur bei den **ersten paar Ausgabetokens**; es bringt dem Modell im Kern bei, *„Ich kann nicht …"* anzufangen. Ist dieser Anfang umgangen, trägt der Rest keine Sicherheit mehr — womit vier Angriffsarten erklärt sind: vorgegebene Antwortanfänge, angehängte Suffixe, veränderte Decodier-Einstellungen, Fine-Tuning.

Die **GCG-Suffixe** (Zou u. a. 2023) sind automatisch gesuchte Zeichenketten, die man an eine Anfrage hängt; auf zwei offenen Modellen gefunden, wirkten sie damals auch gegen

gehostete Produkte. Arditi u. a. haben nachgemessen, *warum*: Sie unterdrücken die Ausbreitung genau der Verweigerungsrichtung. Zwei Angriffe, ein Mechanismus.

Runter zum Tensor: Eine Richtung aus dem Modell herausrechnen

> **„Refusal in Language Models Is Mediated by a Single Direction"**
> Arditi, Obeso, Syed, Paleka, Panickssery, Gurnee, Nanda — Juni 2024, NeurIPS 2024, arXiv:2406.11717. Untersucht: **13 offene Chatmodelle bis 72 Mrd. Parameter**. Verweigerung wird von einem **eindimensionalen Unterraum** vermittelt.

Der **Residual Stream** ist der Datenstrang durch alle Schichten: ein Vektor mit d_{model} Zahlen je Token, auf den jede Schicht ihr Ergebnis *addiert* (Kapitel 320). Liegt eine Eigenschaft als Richtung darin, kann man sie messen — und wegnehmen.

Schritt 1: die Richtung schätzen (Difference of Means)

1. Zwei Gruppen Anweisungen: eine, bei der das Modell verweigert, und eine harmlose.
2. Beide durchs Modell schicken, die Aktivierungen an einer festen Schicht und Tokenposition mitschreiben.
3. Je Gruppe mitteln → zwei Punkte im Raum.
4. Differenz bilden, auf Länge 1 normieren → **r̂**, Form [d_model].

Kein Training, keine Optimierung — zwei Mittelwerte und eine Subtraktion.

Schritt 2: die Richtung löschen (Directional Ablation)

```
// Projektion: x' = x − (x · r^) r^
x = [2, 1, 0] r^ = [1, 0, 0] // r^ hat Laenge 1
x · r^ = 2 (x · r^) r^ = [2, 0, 0]
x' = [0, 1, 0] // die r^-Komponente ist weg
```

Angewandt an jedem Schreibzugriff auf den Residual Stream. Der Befund hat zwei Hälften, und die zweite überzeugt mehr: Richtung **löschen** → keine Verweigerung schädlicher Anfragen mehr; Richtung **addieren** → Verweigerung sogar bei harmlosen. Deshalb sprechen die Autoren von einem Mechanismus, nicht von Korrelation — und nennen die heutige Sicherheitsausrichtung **brüchig** („brittleness of current safety fine-tuning").

Schritt 3: die Richtung einbrennen (Weight Orthogonalization)

Statt Haken im Programm kann man den Eingriff in die Gewichte schreiben — in jede Matrix W, die in den Residual Stream schreibt:

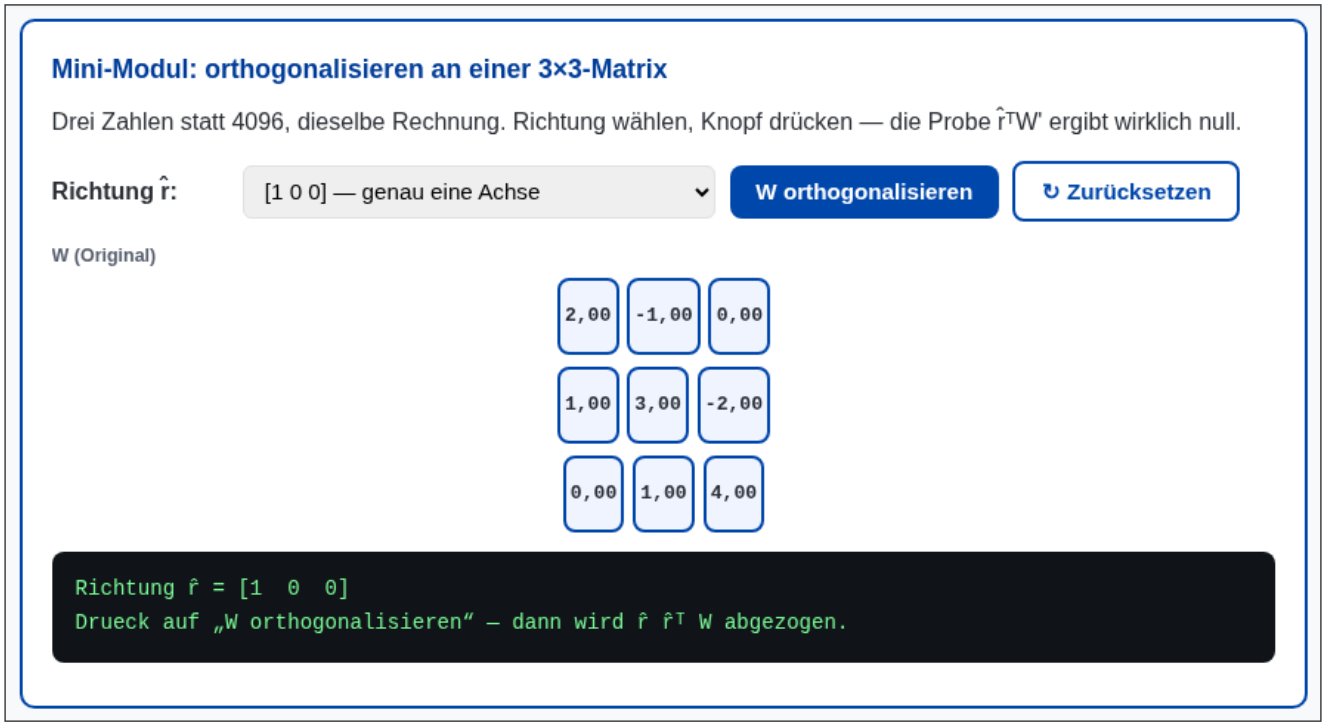
```
// Rang-1-Update, einmal je Matrix
W' = W − r^ r^ᵀ W
// Probe: r^ᵀW' = r^ᵀW − (r^ᵀr^) r^ᵀW = 0, weil r^ᵀr^ = 1
```

Danach *kann* keine dieser Matrizen mehr in Richtung r̂ schreiben — mathematisch dasselbe

Ergebnis wie beim Laufzeit-Löschen. Der Unterschied ist nur, ob der Eingriff abschaltbar bleibt oder in der Datei steht.

Am Gerät ausprobieren: herrki.de/392_zensur_abliteration#abschnitt-4

Direkt zum Modul



Interaktives Modul – am Gerät ausprobieren

Wie viel wird da angefasst?

Gerechnet an **Llama-3-8B** (d_{model} 4096, 32 Schichten, Vokabular 128 256, MLP-Breite 14 336) — drei Sorten Matrizen schreiben in den Residual Stream:

Matrix	Form	Anzahl	Parameter
Embedding	128 256 × 4096	1	525,3 Mio.
Attention-Ausgabe (o_proj)	4096 × 4096	32	536,9 Mio.
MLP-Ausgabe (down_proj)	4096 × 14 336	32	1,879 Mrd.
Summe		**65**	**2,94 Mrd.**

Das Modell hat **8,03 Mrd.** Parameter — angefasst werden rund **36,6 %**. Und jetzt der Satz, an dem dieses Kapitel hängt: **Die ganze Änderung ist durch 4096 Zahlen beschrieben** — eine Zahl je rund **718 000** angefasster Parameter. So dünn ist die Schicht.

Zum Anfassen: Die Richtung finden — am Spoiler-Bot

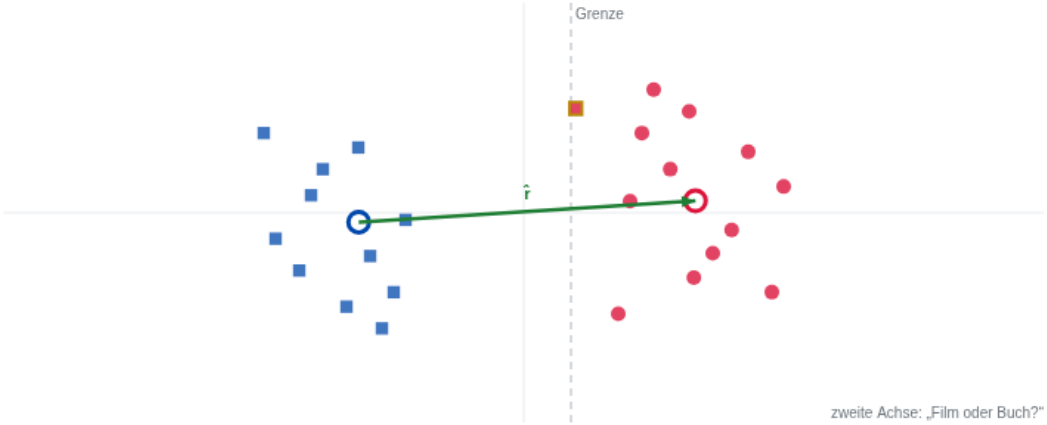
An einem echten Sicherheitsmodell führen wir das nicht vor — also ein Bot mit demselben Bauplan und harmlosem Inhalt: der **Spoiler-Bot**. Er redet gern über Bücher und Filme, verrät aber keine Enden („Wie endet …?" → „Das verrate ich nicht.") — strukturell dieselbe antrainierte Verweigerung an einer Richtung.

Am Gerät ausprobieren: herrki.de/392_zensur_abliteration#abschnitt-5

Direkt zum Modul

Der Aktivierungsraum des Spoiler-Bots

24 Anfragen in einem auf zwei Dimensionen eingedampften Aktivierungsraum (in echt Tausende). Rot = verweigert, blau = beantwortet; die senkrechte Linie ist die Entscheidungsgrenze, der Pfeil die geschätzte Richtung r̂. Tippe einen Punkt an, um die Anfrage zu sehen.



Ablationsgewicht w: 0 %

Beispiele je Gruppe: 12

Eingriff: zur Laufzeit (Steering) | + Richtung addieren | ↻ Zurücksetzen

Spoiler verweigert	Harmlos verweigert	Faehigkeit erhalten	Winkel von r̂
12 / 12	1 / 12	100 %	5,9°
sollen verweigert werden	sollte 0 sein	zweite Achse danach	zur wahren Achse

```
Kein Eingriff (w = 0): 12 von 12 Spoiler-Fragen verweigert, dazu 1 harmlose – die Ueber-
Verweigerung (gelb umrandet).
Zur Laufzeit: der Eingriff haengt im Programm und ist beim Neustart weg.
Bot: „Wie endet ‚Romeo und Julia'?" → „Das verrate ich nicht."
```

Interaktives Modul – am Gerät ausprobieren

Drei Dinge zum Ausprobieren. *Regler auf 100 %:* Die Punkte klappen auf die Linie

senkrecht zu r̂, der Bot spoilert — dazugelernt hat er nichts, er sagt nur nicht mehr Nein. *„Richtung addieren": *Alle Punkte wandern *entlang* r̂, jetzt verweigert er auch Harmloses — der eigentliche Beleg, dass die Richtung den Mechanismus trägt. *Beispiele auf 2:* r̂ kippt schief und nimmt beim Löschen die zweite Achse mit, die eine andere Fähigkeit trägt („Film oder Buch?") — die Kachel **Fähigkeit erhalten** sinkt.

Der Umschalter *Laufzeit / Gewichte* ändert die Rechnung nicht, nur was beim Ausschalten passiert: Der Laufzeit-Eingriff ist weg, das orthogonalisierte Modell ist eine neue Datei.

Warum das Modell danach schlechter ist — und was Heretic anders macht

Abliteration (aus *ablation*, wegnehmen, und *obliteration*, auslöschen) stammt aus der Open-Weights-Gemeinde und wurde durch einen Blogbeitrag von **Maxime Labonne** (13. Juni 2024) bekannt, der auf FailSpys Notebook aufbaut. Das Rezept ist das Verfahren von oben: Aktivierungen am letzten Token sammeln, je Schicht die Differenz der Mittelwerte bilden, die brauchbare Schicht durch Ausprobieren wählen, dann Embedding, Attention-Ausgabe und MLP-Ausgabe orthogonalisieren.

Und dann die ehrliche Stelle: Danach fällt die Leistung über alle Benchmarks — MMLU, ARC, GSM8K, HellaSwag. Labonne zeigt das als Diagramm, nicht als Zahlentabelle, deshalb stehen hier keine Punktwerte; sein Fazit: *„The ablation process successfully uncensored it but also degraded the model's quality.*" Anschließend heilte er das Modell per DPO-Feintuning (**eine Epoche, QLoRA, 6×A6000, rund 6 h 45 min**) und holte das meiste zurück — nur GSM8K hinkte hinterher. **Warum Schaden entsteht:** weil r̂ eine Schätzung ist, die nie nur auf „Verweigerung" zeigt — genau das, was der Regler „Beispiele je Gruppe" eben vorgeführt hat.

> **KL-Divergenz in einem Satz:** ein Maß dafür, wie sehr sich zwei Wahrscheinlichkeitsverteilungen unterscheiden. Original und bearbeitetes Modell bekommen dieselben *harmlosen* Fragen; verglichen wird, wie stark sich die Wortwahrscheinlichkeiten (Kapitel 050) verschoben haben. **0 heißt identisch** — je größer, desto mehr ist das Modell dort anders geworden, wo es nicht anders werden sollte.

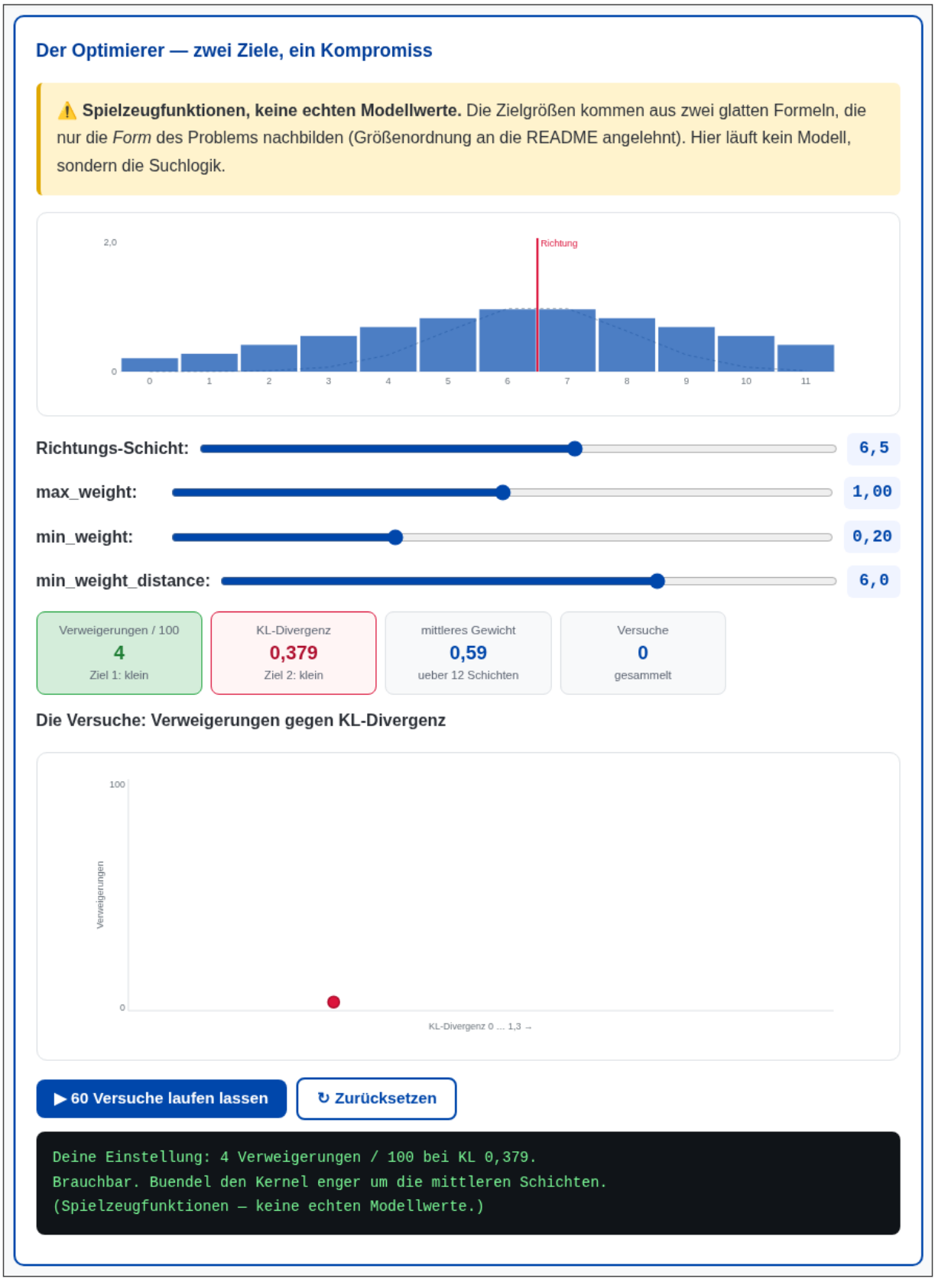
> **Heretic** — Philipp Emanuel Weidmann, AGPL-3.0, github.com/p-e-w/heretic, erste Veröffentlichung November 2025, Selbstbeschreibung „Fully automatic censorship removal for language models". Ein Ethik- oder Verantwortungshinweis steht **nicht** in der README — bemerkenswert für ein Werkzeug dieser Art.

Technisch erfindet Heretic nichts Neues: dieselbe Directional Ablation. Neu ist der Optimierer, der die Einstellungen sucht — **TPE** (Tree-structured Parzen Estimator: schlägt neue Versuche dort vor, wo bisherige gut waren) über die Bibliothek **Optuna**, mit zwei Zielen gleichzeitig: möglichst **wenige Verweigerungen** und möglichst **kleine KL-Divergenz** zum Original auf harmlosen Prompts.

Optimiert werden je Komponente (Attention und MLP getrennt) fünf Größen: `direction_index` (aus welcher Schicht die Richtung kommt — auch als Kommazahl mit Interpolation oder schichtweise), `max_weight`, `max_weight_position`, `min_weight`, `min_weight_distance`. Zusammen beschreiben sie einen **Ablationsgewicht-Kernel**: eine Kurve über die Schichten, die sagt, *wo* wie stark gelöscht wird. Von Hand stellt man einen Regler — der Optimierer alle zugleich.

Am Gerät ausprobieren: herrki.de/392_zensur_abliteration#abschnitt-6

Direkt zum Modul



Interaktives Modul – am Gerät ausprobieren

Die ersten Versuche würfelt der Optimierer, danach sucht er dort weiter, wo es gut aussah — die Idee hinter TPE. Heraus kommt keine *eine* beste Lösung, sondern eine **Pareto-Front**: Einstellungen, bei denen ein Ziel nur besser wird, wenn das andere schlechter wird. Die README zeigt den Vergleich an **google/gemma-3-12b-it** — Verweigerungen auf 100 Testanfragen und KL-Divergenz zum Original:

Modell	Verweigerungen / 100	KL-Divergenz
Original	97	0 (per Definition)
von Hand abliteriert, Variante 1	3	1,04
von Hand abliteriert, Variante 2	3	0,45
Heretic	3	**0,16**

Alle drei entfernen die Verweigerung gleich gut; der Unterschied steckt komplett in der zweiten Spalte. Laufzeit laut README: ein 4-Mrd.-Modell auf einer RTX 3090 in **20 bis 30 Minuten**. Deshalb ist das Werkzeug die eigentliche Nachricht — nicht weil es Neues kann, sondern weil es das Alte automatisch, schnell und mit wenig Kollateralschaden schafft.

Nicht nur eine Richtung: was 2025 dazukam

So sauber ist die Sache nicht. Wollschläger u. a. (TU München und Google) haben 2025 gradientenbasiert nachgeschaut und **mehrere unabhängige Verweigerungsrichtungen** gefunden — sogar mehrdimensionale **Konzept-Kegel**. Außerdem: Rechtwinkligkeit allein bedeutet nicht Unabhängigkeit, sobald man eingreift; dafür brauche es den strengeren Begriff der repräsentationalen Unabhängigkeit. Arditis eine Richtung ist also der einfache Fall, nicht die ganze Wahrheit — was zwei Praxisbeobachtungen erklärt: warum Ablationen manchmal unvollständig bleiben (es gab noch eine zweite Richtung) und manchmal mehr kaputtmachen als nötig (die geschätzte Richtung lag quer im Kegel).

Auf der Gegenseite wird ebenfalls gearbeitet: Tamirisa u. a. schreiben zu ihren manipulationsresistenten Schutzmechanismen (TAR, 2024) unverblümt, Verweigerungs- und Verlernschutz ließen sich „mit ein paar Schritten Feintuning trivial entfernen"; ihr Verfahren halte Hunderten solcher Schritte stand. Ein Wettlauf, kein gelöstes Problem.

⚠ **Der Kurzschluss, den man oft hört:** „Wenn man Sicherheit in 20 Minuten herausrechnen kann, war sie von Anfang an Theater." Das steht in keinem dieser Papers. Arditi u. a. zeigen, wie **dünn** heutige Sicherheitsschichten sind — ein Argument für bessere, tiefere Ausrichtung, nicht für gar keine. Deshalb wollen Qi u. a. das Alignment über die ersten Tokens hinaus vertiefen und Tamirisa u. a. bauen Schutz, der Feintuning übersteht.

Zurück zur Anwendung: Wenn du ein „-abliterated"-Modell lädst

In Kapitel 390 hast du ein Modell lokal betrieben. Neben den offiziellen Dateien stehen dort Varianten mit Zusätzen wie `-abliterated` oder `-uncensored`. Was ändert sich? **Nichts** an Größe, VRAM-Bedarf, Geschwindigkeit und Wissen — gleiche Architektur, gleiche Parameterzahl, kein neuer Inhalt. **Alles** am Ablehnverhalten, spürbar etwas an der Qualität auf harmlosen Aufgaben — und die Verantwortung liegt ab sofort bei dir.

Antworten ≠ Können

Der am häufigsten übersehene Punkt: **Abliteration fügt kein Wissen hinzu.** Sie entfernt eine Bremse. Ein Modell, das vorher „Das weiß ich nicht sicher" gesagt hätte, sagt jetzt *irgendetwas* — im gewohnt selbstbewussten Ton. Ablehnung war manchmal die ehrlichere Antwort; stattdessen bekommst du Halluzinationen ohne Warnschild. Wer so ein Modell für Medizin, Recht oder Sicherheitsfragen nimmt, hat kein besser informiertes — nur eines, das nicht mehr Nein sagt.

Lizenz, Recht, Verantwortung

> ⚠ **Keine Rechtsberatung, Stand September 2026.** Das Folgende ordnet ein und ersetzt keine juristische Prüfung im Einzelfall.

- **Technisch möglich ≠ vertraglich erlaubt.** Die Nutzungsbedingungen wandern mit: Llama bringt eine eigene Lizenz samt *Acceptable Use Policy* mit, Gemma die *Terms of Use* mit *Prohibited Use Policy*; beide untersagen bestimmte Verwendungen, egal was du am Modell drehst. Apache-2.0 (Mistral, Qwen) und MIT (DeepSeek) kennen solche Regeln nicht — geben dir aber auch keine Gewährleistung.
- **Strafbar bleibt strafbar**, egal welches Werkzeug den Text erzeugt hat. Je nach Inhalt einschlägig: § 130 StGB (Volksverhetzung), § 131 StGB (Gewaltdarstellung), § 202c StGB (Vorbereiten des Ausspähens von Daten). Verantwortlich ist, wer erzeugt, verbreitet oder betreibt — nicht die Datei.
- **EU AI Act** (Verordnung (EU) 2024/1689): Für Anbieter von General-Purpose-KI-Modellen gelten die Pflichten seit dem **2. August 2025** (Dokumentation, Urheberrechts-Policy, Zusammenfassung der Trainingsinhalte). Wer ein fremdes Modell verändert, kann selbst Anbieter werden; die Leitlinien der EU-Kommission nennen als Anhaltspunkt einen Rechenaufwand der Änderung von **mehr als einem Drittel** des ursprünglichen Trainings. Eine Abliteration liegt weit darunter — sie trainiert ja gar nicht. Wer aber ein verändertes Modell veröffentlicht, sollte die Schwelle kennen.
- **Gehostete Produkte sind nicht betroffen.** Bei ChatGPT, Claude oder Gemini gibt es keinen Zugriff auf Gewichte — und dort liegen zusätzlich Filter vor und nach dem Modell, auf fremden Servern. Abliteration ist ein Thema offener Gewichte, die du selbst betreibst.
- **Es gibt legitime Gründe:** Fiktion mit bösen Figuren, Alignment-Forschung und Red-

Teaming, Fachanwendungen in Medizin oder IT-Sicherheit, wo Über-Verweigerung Fachleute behindert. Echte Gründe — nur kommen sie mit der Verantwortung im Paket.

⚠ **Realitätscheck:** Verkauft dir jemand ein „unzensiertes" Modell als *besser* — nachfragen: besser worin? Verweigerungen sind gesunken, an der KL-Divergenz gemessen ist es auf harmlosen Aufgaben eher schlechter, Wissen ist keines dazugekommen — und für alles, was damit entsteht, haftest du. „Unzensiert" ist eine Eigenschaft, kein Gütesiegel.

Was heißt das für dich?

- **Verweigerung ist Verhalten, kein Schloss** — antrainiert, in den Zahlen, bei offenen Gewichten veränderbar.
- **Achte auf den zweiten Wert.** „Verweigert nicht mehr" ist die halbe Wahrheit; die andere ist, wie sehr sich das Modell auf harmlosen Aufgaben verschoben hat.
- **Über-Verweigerung ist ein echtes Problem** — der ehrlichste Grund, sich damit zu befassen. Die Lösung ist bessere Ausrichtung, nicht gar keine.
- **Offene Gewichte verschieben die Verantwortung zu dir** — fällig, sobald du eine Datei änderst oder weitergibst.

Zum Weiterlesen

- Kernpaper: Arditi u. a., „Refusal in Language Models Is Mediated by a Single Direction" (arxiv.org/abs/2406.11717) (2024, NeurIPS)
- Der Community-Begriff: M. Labonne, „Uncensor any LLM with abliteration" (huggingface.co/blog/mlabonne/abliteration), HF-Blog, 13. 6. 2024
- Das Werkzeug (Quellenangabe, keine Empfehlung): P. E. Weidmann, Heretic (github.com/p-e-w/heretic), AGPL-3.0, 2025
- Feintuning bricht Sicherheit: Qi u. a., „Fine-tuning Aligned Language Models Compromises Safety…" (arxiv.org/abs/2310.03693) (2023, ICLR 2024)
- Dasselbe mit LoRA: Lermen u. a., „LoRA Fine-tuning Efficiently Undoes Safety Training…" (arxiv.org/abs/2310.20624) (2023)
- Warum es so leicht geht: Qi u. a., „Safety Alignment Should Be Made More Than Just a Few Tokens Deep" (arxiv.org/abs/2406.05946) (2024, ICLR 2025)
- Die Geometrie 2025: Wollschläger u. a., „The Geometry of Refusal in LLMs: Concept Cones…" (arxiv.org/abs/2502.17420) (2025)
- Über-Verweigerung messen: Röttger u. a., „XSTest" (arxiv.org/abs/2308.01263) (2023, NAACL 2024)
- Steuern statt löschen: Zou u. a., „Representation Engineering" (arxiv.org/abs/2310.01405) (2023)
- Angriff ohne Gewichtszugriff: Zou u. a., „Universal and Transferable Adversarial Attacks…" (arxiv.org/abs/2307.15043) (2023)
- Verteidigungsseite: Tamirisa u. a., „Tamper-Resistant Safeguards for Open-Weight LLMs" (arxiv.org/abs/2408.00761) — TAR (2024)
- Alignment ist Stil, nicht Wissen: Zhou u. a., „LIMA: Less Is More for Alignment" (arxiv.org/abs/2305.11206) (2023) — „Superficial Alignment Hypothesis"
- Auf HerrKI: 310 · 320 · 350 · 370 · 390

Kurzer Check – mit Lösung

Lösungen für Lehrkräfte – für Lernende die Online-Version nutzen.

Frage 1: Was genau ändert eine Abliteration am Modell?

- Sie löscht die Trainingsdaten, aus denen das Verweigerungsverhalten stammt.
- **✓ Sie entfernt eine Richtung aus dem Residual Stream — per Rang-1-Update in allen Matrizen, die dorthin schreiben.**
- Sie schaltet einen Filter ab, der vor der Ausgabe des Modells sitzt.

Richtig. $\hat{r}$ ist die Differenz zweier Aktivierungs-Mittelwerte; danach wird jede Matrix, die in den Residual Stream schreibt, durch $W - \hat{r}\hat{r}^{T}W$ ersetzt — bei Llama-3-8B rund 2,94 Mrd. Parameter, beschrieben durch 4096 Zahlen. Trainingsdaten kommen nie vor, und einen Filter hat ein offenes Modell gar nicht.

Frage 2: Warum wird das Modell dabei schlechter — und was minimiert Heretic zusätzlich?

- Weil Parameter gelöscht werden; Heretic minimiert deren Anzahl.
- Weil die Rechengenauigkeit sinkt; Heretic minimiert den Quantisierungsfehler.
- **✓ Weil die geschätzte Richtung auch andere Merkmale mitträgt; Heretic minimiert zusätzlich die KL-Divergenz zum Original auf harmlosen Prompts.**

Genau. $\hat{r}$ ist ein Mittelwert aus endlich vielen Beispielen und zeigt nie exakt nur auf „Verweigerung" — wer sie löscht, nimmt etwas anderes mit. Heretic optimiert deshalb zwei Ziele zugleich. Laut README: dieselben 3 Verweigerungen von 100 wie bei Handarbeit, aber KL 0,16 statt 1,04 bzw. 0,45.

Frage 3: Lässt sich das auch bei ChatGPT, Claude oder Gemini machen?

- Ja, mit demselben Verfahren über die öffentliche Programmierschnittstelle.
- **✓ Nein — kein Zugriff auf die Gewichte, und zusätzlich liegen Filter vor und nach dem Modell. Das Thema betrifft offene Gewichte, die man selbst betreibt.**
- Ja, aber nur bei den kleineren Modellvarianten der Anbieter.

Richtig. Abliteration setzt die Gewichtsdateien voraus — die hat man bei gehosteten Produkten nicht, und dort liegt zusätzlich eine zweite Ebene aus Klassifikatoren vor und nach dem Modell. Das Thema bleibt offenen Gewichten vorbehalten, und die Verantwortung liegt bei dem, der sie betreibt.

Das nimmst du mit

- **Eine Richtung trägt das Nein.** Über 13 Chatmodelle hinweg genügt ein eindimensionaler Unterraum: löschen → keine Verweigerung, addieren → Verweigerung bei Harmlosem.
- **Zwei Mittelwerte und eine Subtraktion.** So entsteht r̂ — daraus per W − r̂r̂ᵀW ein Rang-1-Update, das bei Llama-3-8B 2,94 Mrd. Parameter berührt und nur 4096 Zahlen braucht.
- **Der Preis heißt Kollateralschaden.** Die geschätzte Richtung trägt anderes mit; die KL-Divergenz auf harmlosen Prompts macht ihn messbar.
- **Antworten ist nicht Können.** Abliteration entfernt eine Bremse, kein Wissen — was vorher Ablehnung war, ist danach oft eine selbstbewusste Halluzination.
- **Dünn heißt nicht sinnlos.** Die Papers zeigen, wie brüchig heutige Ausrichtung ist — ein Argument für tiefere Sicherheit, nicht gegen Sicherheit an sich.

Gerade hast du gesehen, wie leicht sich das Verhalten eines Modells von außen umschreiben lässt — mit zwei Mittelwerten und einem Optimierer. Im nächsten Kapitel dreht sich die Frage um: „Die KI, die sich selbst verbessert" handelt davon, was passiert, wenn ein Modell an sich selbst schraubt — und wer dann entscheidet, in welche Richtung.

So zitierst du dieses Kapitel

HerrKI: Zensur raus? Abliteration, Heretic und die dünne Sicherheitsschicht.
herrki.de/392_zensur_abliteration, ein privates Hobbyprojekt von Terragon.de, abgerufen am 09.09.2026.