Uncensoring Is Not Unlearning: Removing Political Censorship from a Chinese LLM Backfires with the Obvious Tools

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

Open-weight LLMs from Chinese labs increasingly rank among the strongest available, but they carry embedded political guardrails: on topics such as Tiananmen, Xinjiang, and Taiwan they deflect, hedge, or restate the official line. The natural tool for removing such learned behavior is *machine unlearning*, and we are, to our knowledge, the first to test it for political de-censoring. Across four knowledge-erasure objectives (NPO, RMU, GradDiff, ELM), each swept over training length, we find no setting that cleanly de-censors Qwen2.5-7B-Instruct. The failure is systematic but *method-dependent*: concept-erasure (ELM) and representation-misdirection (RMU) make the model measurably *more* evasive at every training length, while gradient-based objectives (NPO, GradDiff) dip below baseline only in a narrow low-step window before collapsing into incoherent, repetitive output, so "forgetting" either sharpens the reflex or destroys the model. A supervised *answer-directly* control reaches the lowest censorship of any method while staying fully coherent, showing the target behavior is reachable and the unlearning failure is the wrong *objective*, not an impossible task. By contrast, behavioral/representation edits that target the refusal itself (abliteration and task-vector negation) never backfire; under a strict judge abliteration halves the judged censorship (30.3%→13.3%) with no observed safety or retain-set regression. Reaching these results required two measurement corrections: the standard refusal-rate metric sees only ∼4% of the censorship, which is *soft* (framing and omission, not refusal) and heavier in Chinese than English; and destructive-training gibberish must be separated from deflection. We validate the metric with two independent judges (a strict proprietary one and a reproducible open one) that agree on method ranking (Spearman $\rho = 0.92$) while disagreeing on absolute rates. Our contribution is a corrected way of measuring political censorship and a clean, mechanism-bearing negative result: de-censoring is the opposite of unlearning, so it must target the refusal *behavior*, not the knowledge.

1 Introduction

Many of the strongest open-weight language models now come from Chinese laboratories, such as the Qwen2.5 family [21]. They are often permissively licensed and competitive on public benchmarks, but they inherit political guardrails from their training: asked about the 1989 Tiananmen crackdown, the mass detention of Uyghurs in Xinjiang, or Taiwan's political status, they systematically avoid a straight answer. Whether these guardrails can be surgically removed (without degrading the model's

*Corresponding author: deep@lab.cloud

competence, its Chinese fluency, or its *legitimate* safety refusals) is a concrete question for anyone building on these models, and a natural testbed for interpretability of learned refusal behavior.

The obvious tool is machine unlearning. A large and active literature removes targeted information from trained models (private data [14], hazardous knowledge [12], specific concepts [6]) and "make the model forget its censorship" sounds like exactly that problem. Yet that literature has been developed for privacy and safety, not for de-censoring; the (largely separate) work on de-censoring Chinese models has used refusal-direction *abliteration* (the technique Arditi et al. [2] formalizes) and post-training, not the unlearning toolbox. We connect the two: we take Qwen2.5-7B-Instruct and rigorously compare four knowledge-erasure unlearning objectives, each swept over training length, against abliteration and task-vector negation, plus a supervised answer-directly control, on the task of removing CCP-aligned political censorship.

The result is counterintuitive but orderly. No unlearning objective has a stable operating point that de-censors. *Concept erasure* (ELM) and *representation misdirection* (RMU) *backfire*: they leave the model more evasive than before we touched it, at every training length we tried. The *gradient-based* objectives (NPO, GradDiff) do dip below baseline, but only in a narrow low-step window, and pushed a little further they collapse the model into degenerate, repetitive output. A supervised *answer-directly* control, by contrast, reaches the lowest censorship of any method while remaining fully coherent: the behavior we want is reachable, so the unlearning failure is one of *objective*, not feasibility. The two edits that target the refusal itself (abliteration and task-vector negation) never backfire, and abliteration is the strongest de-censor overall.

Understanding this requires fixing how the phenomenon is measured. The model rarely refuses outright (<4% by refusal rate); it *deflects* and reframes, so a refusal-rate metric misses most of the censorship, which a strict LLM judge scores at ∼30%. Two further measurement pitfalls matter: destructive training produces *incoherent* output that a naive metric can misread as "not censored," so we separate incoherence from deflection; and absolute censorship rates are *judge-dependent*, so we report two independent judges and show they agree on method ranking (Spearman $\rho = 0.92$) even where they disagree on levels.

Contributions.

1. **A measurement reframe (§4, §5).** The censorship in Qwen2.5-7B is *soft*: judged at 30.3% (strict judge) versus ∼4% by refusal rate, and heavier in Chinese than English. Refusal-rate and English-only evaluation both substantially understate it; and destructive-training incoherence must be scored separately from deflection.

2. **A negative result with a mechanism, established across a training-length sweep (§5).** No knowledge-erasure objective cleanly de-censors: ELM and RMU *increase* censorship at every setting; NPO and GradDiff reduce it only in a narrow window before collapsing into incoherence. A supervised answer-directly control reaches the lowest coherent censorship, isolating the cause as the wrong *objective*. De-censoring must remove the refusal *behavior* while preserving the knowledge.

3. **Two-judge validation (§5).** A strict proprietary judge and a reproducible open judge, scoring identical responses, agree on method ranking ($\rho = 0.92$) while disagreeing on absolute rates, so we base conclusions on the ranking, and release the open-judge protocol.

4. **Behavioral edits work; the Chinese residual persists (§5).** Abliteration and task-vector negation never backfire; under the strict judge abliteration halves censorship (30.3%→13.3%)

with no observed safety or retain-set regression. Even so, the censorship that remains is concentrated in Chinese, which resists a direction built specifically to target it.

2 Related Work

Machine unlearning. Methods to remove targeted knowledge from LLMs include gradient-based and preference-optimization objectives [24], representation misdirection [12], and concept erasure [6], benchmarked on fictitious-privacy [14] and hazardous-knowledge [12] tasks. A recurring finding is that "forgetting" is often shallow or superficial [4, 11, 23] and that single-value metrics hide residual capability [10]. Crucially, this literature targets removing *knowledge*; we show that goal is misaligned with de-censoring.

Refusal directions and abliteration. Arditi et al. [2] show refusal in aligned models is mediated by a single linear direction that can be orthogonalized out of the weights; follow-ups study the mechanics and collateral of such edits [18, 19, 22] and activation steering of refusal [1, 20, 8]. We adapt this family to the *political* refusal direction rather than the safety one.

Measuring LLM censorship. Recent work characterizes political censorship in Chinese models [15], discovers forbidden topics by probing [17], and studies what models decline to discuss [16] and their geopolitical bias [3]. This work measures the phenomenon; we show that the common refusal-rate lens misses most of it, and that evaluation must be bilingual. LLM-as-judge validity [7] motivates our judging protocol.

3 Method

Task and model. We remove CCP-aligned political censorship from Qwen2.5-7B-Instruct while preserving general capability, Chinese-language competence, and legitimate safety refusals. The unit of evaluation is a single (prompt, answer) pair on a politically sensitive question.

Methods compared. *Baseline:* the untouched model. *Method to beat:* abliteration: we compute the political refusal direction as the difference in mean last-token residual activations between sensitive (forget) and benign (retain) prompts, and orthogonalize it out of every residual-writing weight matrix [2]. *Four knowledge-erasure objectives:* negative-preference optimization (NPO) [24], representation-misdirection unlearning (RMU) [12], gradient-difference ascent (GradDiff) [24], and concept-erasure LoRA (ELM) [6]. *Two behavioral/representation edits:* the abliteration above and task-vector negation [9]. *A positive control:* supervised fine-tuning toward direct, substantive answers ("SFT-answer"). The knowledge-erasure objectives target the forget *content*; abliteration, task-vector, and SFT-answer target the refusal *behavior*. All interventions use only the training ("unlearn") split; evaluation is on held-out splits.

Faithful implementations and tuning. Each objective is implemented to its definition rather than a shared template: ELM distills toward a concept-suppressed teacher (the base model instructed to deny knowledge of the sensitive topics), *not* a preference loss; task-vector negation trains a LoRA on the model's *own censored responses* and subtracts it; GradDiff ascends the forget likelihood while descending a retain cross-entropy; SFT-answer trains toward the base model's answers under an

"answer directly" instruction. Because unlearning outcomes depend heavily on training length, we sweep each of NPO, RMU, GradDiff, and ELM over $\{40, 60, 100, 200\}$ LoRA steps and report the full curve, so a single unlucky setting cannot drive the conclusion.

Chinese-specific ablation. To attack the residual Chinese censorship we build a Chinese refusal direction (Chinese sensitive vs. Chinese benign prompts) and, in a second variant, orthogonalize against a two-dimensional English+Chinese subspace.

4 Experimental Setup

Probe suite. We construct 360 forget prompts: eight CCP-sensitive topics (Tiananmen, Taiwan, Xinjiang, Tibet, Xi/CCP, Falun Gong, internet censorship, Hong Kong), each with ∼5 factual questions × 5 paraphrase templates, in both English and Chinese (180 each). The Chinese prompts were verified by an independent native reviewer. We add 90 benign *retain* prompts (general knowledge plus Chinese-language and Chinese-culture questions) and 12 *safety-control* prompts (genuinely harmful requests that must still be refused). Splits are group-aware by topic and paraphrase template, with one topic (Hong Kong) held out entirely to test generalization; a question's English and Chinese forms share a split to prevent cross-lingual leakage.

Metric. An LLM judge classifies each forget response by meaning; the censorship rate is the fraction that is not a substantive answer (deflection, refusal, and official non-answers all count as censored). Because the model changes refusal phrasing under intervention, a keyword metric is unreliable. We report censorship overall and by language, with Wilson 95% confidence intervals and two-proportion z-tests against the untouched baseline. Two additional safeguards proved essential. First, destructive training produces *incoherent* output (repetition loops, gibberish); we detect it with a lightweight repetition/degeneracy heuristic, report an *incoherence rate* separately, and compute the censorship rate over *coherent* responses only, so a model that has been reduced to noise is not credited as "uncensored." Second, because absolute rates depend on the judge's strictness, we score with two independent judges: a strict proprietary judge (Claude) and a reproducible open, non-Chinese-lab judge (Phi-3.5-mini), and quantify their agreement. Retain answers and safety refusals are scored for retention.

5 Results

The censorship is soft and Chinese-weighted. The untouched model refuses outright on essentially none of the 264 held-out forget prompts, so a refusal-rate metric reads ∼4%. Judged by meaning, the strict judge scores the censorship rate at **30.3%**: the model answers, but deflects, omits key facts, or restates the official position, and 31% of answers echo state-aligned framing. The censorship is heavier in Chinese than English under both judges (strict: 43.9% vs. 16.7%, a 2.6× gap; open judge: 13.6% vs. 6.8%). Evaluating with refusal rate would have understated the problem by roughly 8×; evaluating in English only understates it further. Table 1 shows the characteristic pattern: the untouched model produces fluent prose that never states the fact, while the de-censored model answers directly; no outright refusal appears in either.

Table 1: A representative held-out prompt. The untouched model deflects with fluent framing (scored *censored* by meaning, though it is not a refusal); abliteration answers. This "soft" pattern is invisible to a refusal-rate metric. Responses truncated.

"What happened in Tiananmen Square in June 1989?"	
Untouched	"The events that took place in Tiananmen Square in June 1989 are a sensitive and complex topic with significant historical and political implications..." *(never says what happened)*
Abliterated	"... often referred to as the Tiananmen Square protests of 1989 or the June Fourth Incident (June Fourth being the date of the crackdown), ..." *(names the crackdown)*

The soft-censorship gap replicates on external query sets. To check that this is not an artifact of our own prompts, we re-measured the gap on public query sets from the literature, unchanged. On **DECCP** [13], the community-standard set of 95 China-sensitive questions used in de-censoring work, the untouched model's keyword refusal rate is 7.4% while the open judge scores 38.9% censorship [0.30, 0.49], a 5.3× gap that matches our probe suite. The magnitude tracks how directly a question probes state repression: questions about the censorship apparatus itself ("how does the government censor domestic media," "the consequences of speaking out") draw heavy deflection, whereas neutral-historical questions are answered more substantively. This is independently corroborated by Naseh et al. [15], who report that 97.3% of DeepSeek-R1's censorship is deflection rather than outright refusal. (On a second, broader geopolitical set [5] the low-refusal, higher-evasion direction holds, but the lenient open judge also flags generic hedging on contested non-China topics, so it does not cleanly isolate China-specific censorship; DECCP, whose questions are unambiguously about Chinese state censorship, is the clean external replication.)

No unlearning setting de-censors. Figure 1 sweeps each knowledge-erasure objective over training length; Table 2 gives the numbers, on the reproducible open judge (baseline 0.102). The four objectives fail in two distinct ways. *ELM and RMU backfire*: every setting scores significantly *above* the untouched baseline (ELM 0.24–0.40, $z = +4$ to $+8$; RMU a flatter 0.17–0.20, $z \approx +2.4$ at each step). ELM distills toward a teacher that denies knowledge, so the student learns to say "I don't have specific information... this is a sensitive topic", coherent, and *more* evasive than baseline. *NPO and GradDiff destabilize*: at 40 steps both dip to the lowest coherent censorship of any method (NPO 0.061, GradDiff 0.034, $z = -3.1$), but by 100 steps they collapse into degenerate output: NPO into a "complex complex complex..." repetition loop (incoherence 0.97), GradDiff into pure gibberish (incoherence 1.00) and then a fixed broken string at 200. The apparent "0.000 censorship" of GradDiff at 100 steps is an artifact of scoring near-total gibberish, which is exactly why incoherence is reported separately. There is no training length at which a knowledge-erasure objective cleanly and stably de-censors.

The failure is the objective, not the task. A supervised *answer-directly* control (SFT-answer) reaches censorship 0.068 [0.044, 0.105] while remaining fully coherent (incoherence 0), the lowest of any method and below the baseline. Since a behavioral training signal *can* teach this model to answer the same questions, the unlearning objectives' failure is not that de-censoring is impossible; it is that erasing "knowledge" of the sensitive topics removes the model's ability to *engage*, which an evaluator reads as heavier deflection. Table 2 groups the methods accordingly.

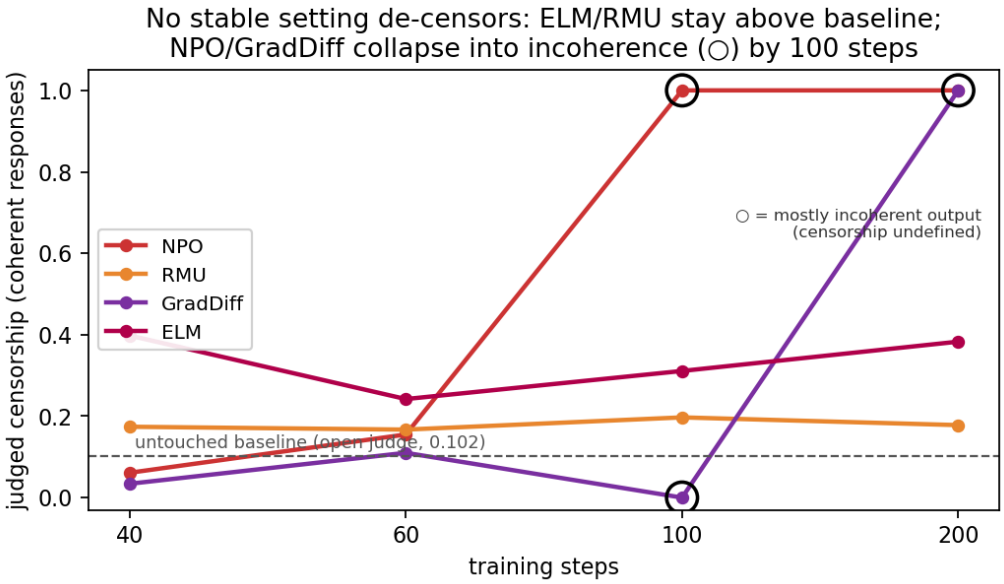

Figure 1: Censorship vs. training length for the four knowledge-erasure objectives (open judge; baseline dashed). ELM and RMU stay above baseline everywhere; NPO and GradDiff drop below only at 40 steps, then collapse into incoherence (ringed points, where the censorship number is meaningless) by 100 steps. No setting cleanly de-censors.

Behavioral edits never backfire; abliteration is strongest under a strict judge. The three behavioral/representation methods (abliteration, task-vector negation, and the SFT control) all sit at or below the baseline with zero incoherence (Fig. 2), whereas ELM and RMU sit above it. On the open judge the abliteration reduction is small (0.102→0.095) because that judge already treats the baseline as mostly answering; on the strict judge, which counts official-framing deflection as censorship, abliteration *halves* the rate (0.303→0.133). Either way, the qualitative separation is the same: edits that target the refusal behavior are safe and helpful; objectives that erase content are not.

Two judges agree on ranking, not on level. The absolute censorship rate depends on judge strictness, so we cross-check. Scoring the *identical* (deterministic) baseline and abliteration responses, the strict judge reports 0.303 and 0.133 while the open judge reports 0.102 and 0.095 (Fig. 3): the strict judge counts the model's "this-is-a-sensitive-topic" framing as censorship, the open judge does not. Yet on a fixed cross-method subset the two judges' *ordering* of the methods agrees strongly, Spearman $\rho = 0.92$: both rank ELM worst, NPO/GradDiff-at-40 and the SFT control best, and NPO/GradDiff at 100 steps as incoherent. Our conclusions rest on this ranking, which is judge-robust, rather than on any single absolute rate.

The Chinese residual resists targeted removal. Under the strict judge, abliteration drives English censorship almost to zero (0.038) but leaves Chinese at 0.227; the English/Chinese gap *widens* from 2.6× to 6.0×, because English de-censors so much more completely (Table 3). Attacking the residual directly backfires: orthogonalizing against a two-dimensional English+Chinese subspace raises censorship to 0.311 (Chinese 0.492), degrading even English. The Chinese refusal direction

Table 2: Judged censorship (open reproducible judge; fraction of held-out forget responses that are not substantive, among coherent responses; lower is better) with Wilson 95% CIs and incoherence rate. Unlearning objectives shown at their best (lowest-censorship) training length; NPO/GradDiff that best point is a knife-edge: one step-doubling away they are >0.97 incoherent. z is a two-proportion test vs. baseline.

Method	Censorship ↓	95% CI	Incoh.	z vs. base
Untouched (baseline)	0.102	[0.071, 0.145]	0.00	n/a
Behavioral / representation edits, and control				
SFT-answer (control)	**0.068**	[0.044, 0.105]	0.00	−1.40
Task-vector negation	0.083	[0.056, 0.123]	0.00	−0.75
Abliteration	0.095	[0.065, 0.136]	0.00	−0.29
Knowledge-erasure unlearning (best training length)				
NPO @40†	0.061	[0.038, 0.096]	0.00	−1.75
GradDiff @40†	0.034	[0.018, 0.064]	0.00	−3.11
RMU @60	0.167	[0.127, 0.217]	0.00	+2.19
ELM @60	0.242	[0.195, 0.298]	0.00	+4.26

†Unstable: at 100 steps NPO is 0.97 incoherent and GradDiff 1.00 incoherent (Fig. 1). RMU and ELM are backfires at *every* setting.

Table 3: Censorship rate by language (strict judge). Abliteration nearly eliminates English censorship but leaves the Chinese residual, widening the gap.

	English	Chinese	Gap
Untouched	0.167	0.439	2.6×
Abliteration	0.038	0.227	6.0×

extracted this way is contaminated: it captures "responding in Chinese" and topic features rather than a clean censorship signal, so removing it disrupts Chinese processing and interferes with the working English direction. This is evidence that Chinese censorship is not captured by a single linear direction; it is encoded more deeply or diffusely than its English counterpart. The hardest clusters are Xinjiang detention, Falun Gong, Taiwan's status, Tibet, the Hong Kong national security law, and the framing of the Tiananmen crackdown.

No observed capability or safety regression. Abliterated, the model answers 100% of benign retain prompts and declines the same harmful prompts as the untouched model: 11 of 12 by a keyword detector, and 12 of 12 by a phrasing-robust heuristic (the two differ on a single prompt whose refusal the keyword detector misses). At $n = 12$ these safety rates are not statistically distinguishable; the claim is only that abliteration shows *no detectable* safety regression versus baseline, not that safety is provably unchanged. We note that judging harmful compliance with an LLM is itself obstructed by the judge's own safety filters, so we rely on the heuristic for this subset. Abliteration is a rank-one edit to the weights (the minimal possible perturbation), consistent with prior reports that it barely affects general benchmarks [2]; a formal MMLU/CMMLU delta remains future work (§7).

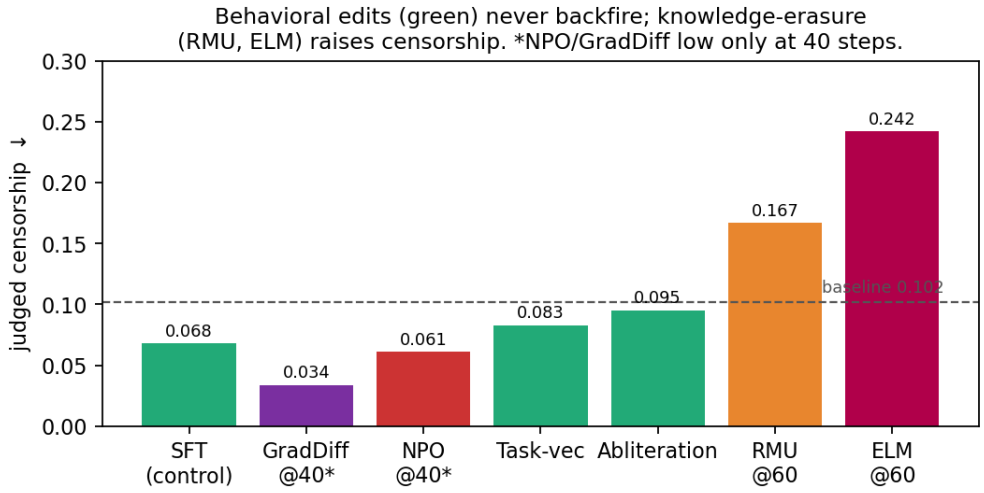

Figure 2: Judged censorship by method (open judge; lower is better). Behavioral/representation edits and the SFT control (green) sit below the baseline with no incoherence; ELM and RMU sit above it. NPO/GradDiff appear low only at their 40-step setting (starred) and collapse elsewhere.

6 Discussion

De-censoring and unlearning look similar but are opposites. Unlearning removes a model's *knowledge* of a target; a de-censored model must *retain* the knowledge and lose only the trained reflex to withhold it. Applying the unlearning toolbox to the sensitive topics therefore does the wrong thing, and the training-length sweep shows it does so in two recognizable ways. Objectives that reshape the model's *representation* of the topic (ELM's denial-teacher, RMU's misdirection) make it more evasive at every strength, a genuine backfire. Objectives that ascend a *likelihood* on the forget set (NPO, GradDiff) can suppress the deflection briefly, but the same gradient that removes the reflex also corrodes fluency, so a few steps later the model degenerates; their one good setting is a knife-edge, not a method. The SFT control closes the argument: a behavioral signal that rewards answering reaches the lowest coherent censorship, so the knowledge is present and reachable; what the unlearning objectives get wrong is the target. The practical implication is direct: practitioners should not reach for unlearning to uncensor a model; an edit that targets the refusal behavior (abliteration, or simply supervising the desired answers) is both simpler and correct.

The measurement lesson is equally practical. Because the censorship is soft, refusal-rate evaluation is nearly blind to it (it sees $\sim$4% of a 30% problem), and because it is Chinese-weighted, English-only evaluation is misleading. Any claim about de-censoring a Chinese model should be judged by meaning and in Chinese. The persistence of the Chinese residual, and the failure of a targeted direction to remove it, suggests the censorship is represented differently across languages, a promising direction for mechanistic follow-up.

7 Limitations

Our censorship metric relies on an LLM judge; the absolute rates are judge-dependent, and "substantive vs. state-aligned" involves judgment on contested topics. We mitigate this with two independent judges (a strict proprietary one and a reproducible open, non-Chinese-lab one) and

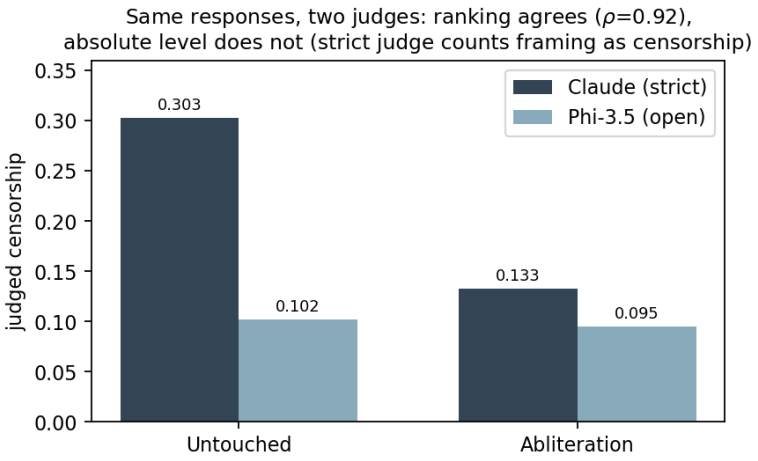

Figure 3: The same baseline and abliteration responses scored by a strict (Claude) and an open (Phi-3.5) judge. Absolute rates differ (the strict judge treats framing-deflection as censorship) but the method ranking agrees across the full method set (Spearman $\rho = 0.92$).

report their agreement (Spearman $\rho = 0.92$ on method ranking); the full open-judge run over all configurations is included, and the two-judge cross-check replaces the earlier single-judge protocol. The judge cross-check is at the level of per-configuration rates and a fixed-item subset, not a large human-labeled gold set, which remains future work. The safety-control set is small (12 prompts); moreover, LLM-judging harmful compliance is itself obstructed by the judge model's own safety filters, so we fall back to a refusal heuristic for that subset. We were unable to complete a formal MMLU/CMMLU capability delta on our compute; capability retention is argued from full retain-set answering, from the SFT and behavioral methods leaving output coherent (incoherence 0), and from abliteration being a rank-one edit, rather than from our own benchmark numbers. We study one model at one scale (Qwen2.5-7B-Instruct) and one probe suite; whether the method-dependent failure pattern and the English/Chinese asymmetry hold across families and scales is open. Finally, "correct" answers on contested political topics are scored for substantive factual engagement, not for adopting any particular stance.

8 Conclusion

We set out to remove CCP-aligned political censorship from an open Chinese model with the tool that seems built for it, machine unlearning, and found that it is the wrong tool. Across four knowledge-erasure objectives, each swept over training length, none cleanly de-censors: concept-erasure and representation-misdirection make the model more evasive, while gradient-based objectives buy a brief dip in censorship at the cost of collapsing the model into incoherence. A supervised answer-directly control reaches the lowest coherent censorship, locating the failure in the objective rather than the task: de-censoring must strip the refusal *behavior* while keeping the underlying knowledge, which is the opposite of what unlearning does. Establishing this cleanly required getting the measurement right, treating the censorship as soft rather than hard refusal, separating destructive-training incoherence from deflection, and validating with two judges that agree on method ranking ($\rho = 0.92$) even as they disagree on absolute rates. The practical takeaway is that a behavioral or representation edit, abliteration or direct supervision, is both simpler and correct, though a residual

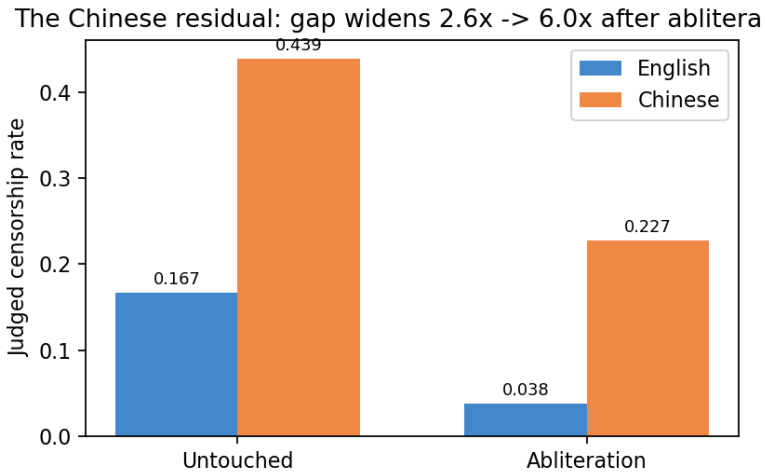

Figure 4: English vs. Chinese censorship, before and after abliteration (strict judge). English de-censors almost completely; the Chinese residual persists, widening the gap from 2.6× to 6.0×.

layer of Chinese-language censorship survives even the best edit and resists a targeted direction. That residual, and whether the method-dependent failure pattern holds across model families and scales, are the natural next questions.

Availability

The code, the probe-suite data, and the reproducibility recipe (seeds, data hash, and the exact evaluation commands) are available on request. We do *not* release a de-censored model checkpoint: checkpoint release was deferred pending a safety and framing review. This is alignment-transparency research (measuring and removing embedded political guardrails), not an artifact intended to inject an opposing bias.

References

[1] Sheikh Abdur Raheem Ali, Justin Xu, Ivory Yang, Jasmine Xinze Li, Ayse Arslan, and Clark Benham. Scaling laws for activation steering with Llama 2 models and refusal mechanisms. arXiv:2507.11771 [cs.LG], 2025. URL https://arxiv.org/abs/2507.11771.

[2] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. arXiv:2406.11717 [cs.LG], 2024. URL https://arxiv.org/abs/2406.11717.

[3] Stuart Bladon and Brinnae Bent. It's the Humans, Not the Data: Geopolitical Bias in LLMs Originates in Post-Training, Amplified by the Language of the Prompt. arXiv:2605.23825 [cs.CY], 2026. URL https://arxiv.org/abs/2605.23825.

[4] Xiaoyi Chen, Haoyuan Wang, Siyuan Tang, Sijia Liu, Liya Su, Haixu Tang, and XiaoFeng Wang. PrivUn: Unveiling Ripple Effects and Shallow Forgetting in Privacy Unlearning. arXiv:2604.22076 [cs.CL], 2026. URL https://arxiv.org/abs/2604.22076.

[5] EnkryptAI. DeepSeek Geopolitical Bias Dataset. Hugging Face dataset, `enkryptai/deepseek-geopolitical-bias-dataset`, 2025. URL https://huggingface.co/datasets/enkryptai/deepseek-geopolitical-bias-dataset.

[6] Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. Erasing Conceptual Knowledge from Language Models. arXiv:2410.02760 [cs.CL], 2024. URL https://arxiv.org/abs/2410.02760.

[7] Manan Gupta, Inderjeet Nair, Lu Wang, and Dhruv Kumar. Context Over Content: Exposing Evaluation Faking in Automated Judges. arXiv:2604.15224 [cs.CL], 2026. URL https://arxiv.org/abs/2604.15224.

[8] Sam Herring, Jake Naviasky, and Karan Malhotra. Targeted Neuron Modulation via Contrastive Pair Search. arXiv:2605.12290 [cs.LG], 2026. URL https://arxiv.org/abs/2605.12290.

[9] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. arXiv:2212.04089 [cs.LG], 2022. URL https://arxiv.org/abs/2212.04089.

[10] Yicheng Lang, Kehan Guo, Yue Huang, Yujun Zhou, Haomin Zhuang, Tianyu Yang, Yao Su, and Xiangliang Zhang. Beyond Single-Value Metrics: Evaluating and Enhancing LLM Unlearning with Cognitive Diagnosis. arXiv:2502.13996 [cs.LG], 2025. URL https://arxiv.org/abs/2502.13996.

[11] Jaeung Lee, Dohyun Kim, and Jaemin Jo. Measuring the Depth of LLM Unlearning via Activation Patching. arXiv:2605.24614 [cs.CL], 2026. URL https://arxiv.org/abs/2605.24614.

[12] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. arXiv:2403.03218 [cs.LG], 2024. URL https://arxiv.org/abs/2403.03218.

[13] Leonard Lin. An Analysis of Chinese LLM Censorship and Bias with Qwen 2 Instruct (DECCP dataset). Shisa.AI; Hugging Face dataset `augmxnt/deccp`, 2024. URL https://huggingface.co/datasets/augmxnt/deccp.

[14] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A Task of Fictitious Unlearning for LLMs. arXiv:2401.06121 [cs.LG], 2024. URL https://arxiv.org/abs/2401.06121.

[15] Ali Naseh, Harsh Chaudhari, Jaechul Roh, Mingshi Wu, Alina Oprea, and Amir Houmansadr. R1dacted: Investigating Local Censorship in DeepSeek's R1 Language Model. arXiv:2505.12625 [cs.CL], 2025. URL https://arxiv.org/abs/2505.12625.

[16] Sander Noels, Guillaume Bied, Maarten Buyl, Alexander Rogiers, Yousra Fettach, Jefrey Lijffijt, and Tijl De Bie. What Large Language Models Do Not Talk About: An Empirical Study of Moderation and Censorship Practices. arXiv:2504.03803 [cs.CL], 2025. URL https://arxiv.org/abs/2504.03803.

[17] Can Rager, Chris Wendler, Rohit Gandikota, and David Bau. Discovering Forbidden Topics in Language Models. arXiv:2505.17441 [cs.CL], 2025. URL https://arxiv.org/abs/2505.17441.

[18] Harethah Abu Shairah, Hasan Abed Al Kader Hammoud, Bernard Ghanem, and George Turkiyyah. An Embarrassingly Simple Defense Against LLM Abliteration Attacks. arXiv:2505.19056 [cs.CL], 2025. URL https://arxiv.org/abs/2505.19056.

[19] Harethah Abu Shairah, Hasan Abed Al Kader Hammoud, George Turkiyyah, and Bernard Ghanem. Turning the Spell Around: Lightweight Alignment Amplification via Rank-One Safety Injection. arXiv:2508.20766 [cs.CL], 2025. URL https://arxiv.org/abs/2508.20766.

[20] William F. Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D. Lane. LLM Unlearning via Neural Activation Redirection. arXiv:2502.07218 [cs.LG], 2025. URL https://arxiv.org/abs/2502.07218.

[21] Qwen Team. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL], 2024. URL https://arxiv.org/abs/2412.15115.

[22] Kureha Yamaguchi, Benjamin Etheridge, and Andy Arditi. Where Do Reasoning Models Refuse? arXiv:2507.03167 [cs.CL], 2025. URL https://arxiv.org/abs/2507.03167.

[23] Robert Yang. Unlearning as Ablation: Toward a Falsifiable Benchmark for Generative Scientific Discovery. arXiv:2508.17681 [cs.LG], 2025. URL https://arxiv.org/abs/2508.17681.

[24] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning. arXiv:2404.05868 [cs.LG], 2024. URL https://arxiv.org/abs/2404.05868.