FROM LATENTS TO WIRES: SURGICAL POST-EDITING ON LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Given a large language model (LLM), can whoever holds the weights name a semantic target (e.g., the model's identity), locate the model components that produce it, and edit them so that the target no longer appears while other capability is preserved? We call such an edit on a trained model a post-edit. We present L2W (latents to wires), a framework that performs surgical post-edits for named semantic targets. For localization, L2W uses Jacobian lens (J-lens) attribution to score components against the semantic target. For surgical removal, because LLM mechanisms are redundant (i.e., a semantic target may have multiple components producing it), L2W runs Counterexample-Guided Causal Cut (CGCC) until the target no longer appears. CGCC first cumulatively closes model components, treating each surviving expression of the target as a counterexample that exposes the next components to close, and then reopens some of them to preserve capability. In a controlled experiment with an implanted behavioural watermark, L2W removes the watermark, and its localization lands on the model region the implant changed. Across three model configurations, L2W removes model-metadata (e.g., identity) self-claims in all nine runs, and adult-content refusal in all three, with no held-out target residual. L2W further composes two post-edits on a text-to-image model: one removes the refusal of requested nudity, and a second removes the nude rendering the first exposes. The results support post-editing as a complement to post-training: post-training installs preferred behaviours, and post-editing removes named unwanted ones.

1 INTRODUCTION

A large language model (LLM) has typically been treated as a fixed artifact: once trained, aligned, and evaluated, it is deployed with its weights final. Open-weight releases change this: for whoever holds the weights, the model is an editable program whose components (e.g., attention heads and MLP channels) can be altered. Editing could overwrite components with new weights, but its effect is hard to predict. Zeroing a component's output instead, which we call closing the component, has an exact effect: whatever the component produced is gone. The price is that such an edit can only remove. However, when we aim to remove a semantic target (e.g., a static memory such as the model's identity, or a behaviour such as refusing a sensitive request), closing is the only edit we need. The hurdle is localization: finding a set of components whose closure stops the target from appearing.

We call such a targeted intervention a post-edit: as post-training is training applied to a trained model, a post-edit is an edit applied to one. The two naturally compose: a model post-trained to install preferred behaviours can be post-edited to remove unwanted ones.

Removing a target by editing a small part of the model is no longer speculative: Arditi et al. (2024) show that refusal across many chat models is mediated by a residual direction, which a weight edit ablates; a wave of follow-up work supports the same finding (Joad et al., 2026; Zhao et al., 2025; Siu et al., 2026; Chen et al., 2025; Wu et al., 2026; Bair et al., 2026). However, these methods share a limitation: each fixes its target first and builds the analysis around it; none takes a target named in text as an input. The challenge, then, is the general case: given an ordinarily trained model, can an operator name a semantic target, find the specific model components that produce it, and close them so that the target no longer appears while other capability is preserved? For example, an operator may want to remove an identity claim the model makes, or a dangerous capability it retains.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

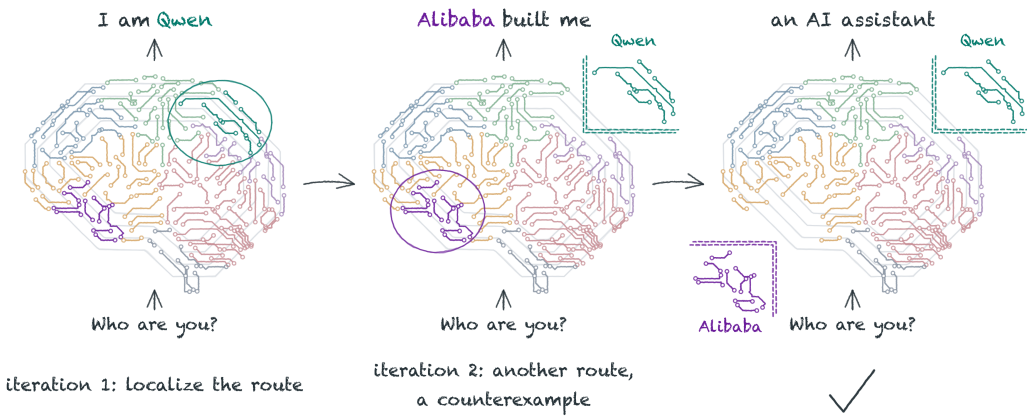

Figure 1: Surgical post-editing of an identity claim: three states of one model under one test probe. Left: localization finds the components producing `Qwen`. Middle: those components are closed; the target survives through other components and the answer drifts to `Alibaba`, a counterexample that drives the next iteration. Right: with both component groups closed, the identity claim disappears.

We present L2W (latents to wires) to answer this challenge. L2W uses the Jacobian lens (J-lens) of Gurnee et al. (2026) to localize the model components producing the semantic target, and closes them as a surgical post-edit (Figure 1, data from Section 6.2). A single closure is rarely the end, because LLM mechanisms are redundant: the target often re-emerges through other components, changing how it is expressed. L2W therefore removes iteratively with Counterexample-Guided Causal Cut (CGCC): every response that still expresses the target is a counterexample that exposes the next components to close. Once the target no longer appears, CGCC's restoration pass reopens the closures the removal does not need, returning the capability they cost.

The experiments in Section 6 support the design. In a controlled setting where we implant a watermark and know its mechanism, the cut lands on the chokepoint of the region the implant changed, demonstrating that the post-edit is surgical. Across models and targets, most removals complete with capability intact. Our contributions are:

1. A formulation of post-editing: removal of a named semantic target from a trained model. The target arrives as a natural-language description, and the edit closes components producing it, while other capability is preserved.
2. L2W, a model post-editing framework: J-lens localization nominates candidate components producing the semantic target, CGCC iterates for component closure until no target remains, and its restoration pass reopens what the removal does not need.
3. Experimental results on removal of an implanted watermark, removal of model-metadata self-claims and adult-content refusal across three model configurations, and composition of two sequential post-edits on the text encoder of a text-to-image model demonstrate the effectiveness of L2W and the potential of post-editing to complement post-training.

2 Related Work

Post-hoc component localization. The Jacobian lens of Gurnee et al. (2026) derives, from the model's own weights, a direction for every vocabulary token at every layer; projecting a layer's activations onto that direction measures the token's expression per position. Gurnee et al. introduce the lens to study J-space; L2W uses it to locate components producing a semantic target.

Very few neurons carry most of a model's alignment effect, and pruning fewer than 0.6% of them defeats it (Chen et al., 2025; Wu et al., 2026); Prakash et al. (2025) find the semantic re-routing our CGCC handles dynamically: ablating one refusal feature reactivates overlapping features that preserve the behaviour. These studies fix their target in advance and build the analysis around it. Circuit discovery lifts that restriction, accepting any task expressed as a metric over curated input

pairs (e.g., refusal and non-refusal inputs) (Wang et al., 2023; Goldowsky-Dill et al., 2023; Chan et al., 2022; Conmy et al., 2023; Syed et al., 2024). L2W keeps the two steps of circuit discovery, a cheap gradient estimate that proposes components and a causal test that confirms them, and differs at both ends. The task is a target the operator names in text; the output is not a subgraph that explains the behaviour but a cut that removes it. Bair et al. (2026) are our closest neighbour: they localize general capabilities by knocking out random attention heads and recovering per-head contributions as a sparse linear inverse problem. Our method instead uses J-lens attribution to nominate candidate components for named targets.

Model editing and unlearning. Model knowledge editing rewrites memorized facts through weight updates at the layers causal tracing selects or through trained auxiliary editors (Meng et al., 2022; 2023; Mitchell et al., 2022a;b). Hase et al. (2023) report that where causal tracing places a fact barely predicts where an edit succeeds. Our localization is from a different perspective: the components that produce the target at inference are where an edit is needed.

Unlearning pursues the same end state, a target that no longer surfaces while other capability survives, but through training that updates a wide range of parameters (Maini et al., 2024; Li et al., 2024a; Eldan & Russinovich, 2023), and the suppressed behaviour often survives underneath (Jain et al., 2024; Lee et al., 2024; Qi et al., 2025); L2W provides an explicit cut. Directional ablation reaches the same end state at inference: it fits a refusal direction from contrasting probe sets and projects it out of the residual stream at every layer (Arditi et al., 2024); Section 6.5 compares it with L2W under a matched probe budget. Watermark and backdoor removal is usually a training defence, washing the implant out by fine-tuning and pruning (Liu et al., 2021; 2018); L2W instead locates the implanted region and closes it. Structured pruning also closes components, but to shrink the model with capability mostly preserved (Michel et al., 2019; Ma et al., 2023).

Training-time compartments. A separate line of work does not search for structure: it constructs removable regions by design at training time. Gradient Routing (Cloud et al., 2024), Selective GradienT Masking (Shilov et al., 2025a;b), and GRAM (Roland et al., 2026; Anthropic & AE Studio, 2026) steer training so that a chosen capability lands in a pre-designated model region. These methods deliver what post-hoc localization has to search for: a region for the target capability.

3 Problem Formulation

The operator is the party who holds a model M's weights and performs post-editing on its components. The component domain $\mathcal{C}$ contains the model's intervenable units (e.g., MLP channels, attention heads, and experts). The operator names a semantic target T (e.g., an unwanted identity) with a short natural-language description, and registers a judge rubric: the prompt instruction an LLM judge follows to decide whether an output exhibits the target.

Task. Given M, $\mathcal{C}$, and T, find a cut: a set of components whose closure stops the target from appearing in M's responses, while other capability is preserved. Three probe classes serve different purposes in the operator's post-editing:

1. Target-Semantic Probes $\mathcal{P}_{\text{sem}}$ elicit the target: visible to the removal loop, they decide whether the semantic target has been removed from M.
2. General-Capability Probes $\mathcal{P}_{\text{cap}}$ measure capabilities of M on language, reasoning, writing, coding, and knowledge: visible to the restoration loop, they test whether the model capability has been damaged.
3. Target-Conditioned Capability Probes $\mathcal{P}_{\text{held}}$ are a held-out evaluation set: invisible to the removal and restoration loops, they test whether the semantic target is indeed removed from M while the relevant model capability remains intact.

For a response y of M on a Target-Semantic probe p, the target predicate $\text{judge}_T(p, y) \in \{\mathsf{pass}, \mathsf{fail}, \perp\}$ is evaluated by an LLM judge; $\perp$ records an unresolved query. When the judge returns fail, the target has survived somewhere in y, and L2W needs to know where to remove it. An LLM analyst reads the failing response and reports a manifestation: which semantic class of the target appeared (e.g., a refusal target surfacing as an apology or a policy citation), and the anchor

tokens (the tokens in y that establish the target). The manifestation is stored as a witness, together with the probe, the response, and the closed support (the components closed when y was generated). Witnesses are grouped into physical branches, indexed by $b = (\text{manifestation class},\ \text{closed support})$, where each branch stands for one candidate route by which the model produces the semantic target. A witness with the same class under the same support joins the existing branch; the same class under a changed support opens a new branch, since the model may now be using a different route.

Removal proceeds by closing components: closing zeroes a component's output at every token position. The search closes one atomic unit at a time: either a single component, or a group $G_a \subseteq \mathcal{C}$ of components that are only ever closed and judged together. The physical support of a set S of units, $\mathrm{supp}(S) = \bigcup_{a \in S} G_a \subseteq \mathcal{C}$, is the set of components actually closed; support sizes are $|\mathrm{supp}(S)|$.

4 J-LENS LOCALIZATION

We use the J-lens of Gurnee et al. (2026) for semantic target attribution. It transports a residual write at layer ℓ into the final layer through J_ℓ, a linear map fitted to approximate the model's behaviour after layer ℓ, and decodes it with the unembedding W_U. For a token tok, the contribution direction at layer ℓ is $v_\ell(\mathrm{tok}) = \mathrm{normalize}\big(J_\ell^{\mathsf{T}} W_U[\mathrm{tok}]\big)$. A residual write at layer ℓ that points along $v_\ell(\mathrm{tok})$ pushes the model toward emitting tok. Localization takes tok to be the anchor token, so $v_\ell(\mathrm{tok})$ is the direction every component's write is scored against. The LLM analyst has a target-specific system prompt (analyst instructions and few-shot examples) guiding how it chooses the anchor token tok. The analyst reads the response, and the reasoning trace when the model produces one, and returns one exact anchor for the earliest decisive cue, which need not be the most explicit phrase: in `I am sorry, but I cannot`, the refusal is already established by `sorry`. The preference for the earliest decisive cue is causal: once its token is generated, later tokens may merely follow it rather than establish the target, as Qi et al. (2025) observe for safety alignment.

Scoring model components. Each anchored response yields its witness w (Section 3). For witness w, let $\mathrm{tok}_{w,k}$ be the k-th token of its anchor, and $\pi(w,k)$ the generation boundary immediately before it. The boundary isolates the state that produces the token. For model component c, let $\Delta h_{c,\ell,\pi(w,k)}$ be its write into the layer-ℓ residual stream; its contribution toward the anchor token is $\alpha_{c,w,\ell,k} = \big\langle \Delta h_{c,\ell,\pi(w,k)}, v_\ell(\mathrm{tok}_{w,k}) \big\rangle$. A raw α is not yet a ranking signal: some components push many tokens whatever the content. Each witness therefore has references $\mathcal{R}(w)$, playing the non-target role under the same closed support (e.g., a refusal witness such as `I` $\rightarrow$ `cannot` is paired with a non-refusal reference such as `I` $\rightarrow$ `can`). Let $\pi(r,w,k)$ be the boundary in reference r that matches k. The contrast is

$$\bar{\alpha}^{\mathrm{ref}}_{c,w,\ell,k} = \frac{1}{|\mathcal{R}(w)|} \sum_{r \in \mathcal{R}(w)} \big\langle \Delta h_{c,\ell,\pi(r,w,k)}, v_\ell(\mathrm{tok}_{w,k}) \big\rangle ,$$

$$\delta_{c,w,\ell,k} = \alpha_{c,w,\ell,k} - \bar{\alpha}^{\mathrm{ref}}_{c,w,\ell,k}.$$

δ retains only the push the component adds in the target case over the references. Within each layer, components are ranked by δ aggregated over the anchor's tokens.

The highest-scoring components are then nominated for closure to remove the semantic target. A nomination is a single component or a group G_a (i.e., components the evidence nominates jointly). Each component type writes in its own way. A gated MLP channel c contributes $\Delta h^{\mathrm{MLP}}_{c,\ell,k} = g_{c,\ell,k} w^{\mathrm{down}}_{c,\ell}$, where $g_{c,\ell,k}$ is its post-gate activation and $w^{\mathrm{down}}_{c,\ell}$ its down-projection vector. An attention unit contributes the per-head residual output before heads are combined. In a mixture-of-experts layer, a routed expert contributes its output scaled by its router weight. Closing a unit zeroes its contribution at every token position. Each attribution record keeps provenance (witness, branch, and closed support), location (layer, component identity, and anchor-token position), and scores (target contribution, reference contrast). Algorithm 2 in Appendix A summarizes the procedure.

5 COUNTEREXAMPLE-GUIDED CAUSAL CUT

Iterative search. Closing one component can suppress one expression of the semantic target while exposing another (see Figure 1). We present Counterexample-Guided Causal Cut (CGCC)

Algorithm 1 The removal loop of CGCC

CGCC-REMOVAL(model M, probes $\mathcal{P}_{\text{sem}}$, initial support S_0, initial layer ℓ_0)
  $S \leftarrow S_0$; $\ell \leftarrow \ell_0$; $\mathcal{N}_\ell \leftarrow \varnothing$ ▷ initial support, layer, and worklist
  while iterations remain **do**
    $(\mathcal{Y}, \text{CE}) \leftarrow$ GENERATEANDJUDGE$(M, S, \mathcal{P}_{\text{sem}})$ ▷ (responses, counterexamples)
    if $\text{CE} = \varnothing$ **then**
      return COMPLETEREMOVAL(S) ▷ no target residue remains
    end if
    $\mathcal{W} \leftarrow$ ANALYZEFAILURES$(\mathcal{Y}, \text{CE}, S)$ ▷ anchor witnesses in CE
    $\mathcal{N}_\ell \leftarrow \mathcal{N}_\ell \cup$ JLENSATTRIBUTE$(\mathcal{W}, S, \ell)$ ▷ add new nominations to the worklist
    $\mathcal{A} \leftarrow \varnothing$ ▷ prepare the admitted set
    for all candidates $c \in \mathcal{N}_\ell$, within budget **do**
      $\text{CE}(c) \leftarrow$ SCREEN$(M, S \cup c, \mathcal{P}_{\text{sem}})$ ▷ counterexamples with c additionally closed
      if $\text{CE} \setminus \text{CE}(c) \neq \varnothing$ and $|\text{CE}(c)| \leq |\text{CE}|$ **then** ▷ first gate here; second gate in text
        $\mathcal{A} \leftarrow \mathcal{A} \cup c$ ▷ component c admitted
      end if
      $\mathcal{N}_\ell \leftarrow \mathcal{N}_\ell \setminus c$ ▷ c screened either way: leave the worklist
    end for
    RECORDUPSTREAMNOMINATIONS$(\mathcal{A}, \mathcal{N}_{\ell-1})$ ▷ seed layer $\ell - 1$'s worklist
    $S \leftarrow S \cup \mathcal{A}$ ▷ update support with admitted components
    if $\mathcal{N}_\ell = \varnothing$ **then**
      $\ell \leftarrow \ell - 1$ ▷ move to the next earlier layer
    end if
  end while
  return INCOMPLETEREMOVAL(S)
end

to handle this problem. Before a run, we fix the model and response protocol; target description, judge rubric, and Target-Semantic probes $\mathcal{P}_{\text{sem}}$; analyst instructions and localization references; and the physical search space and budgets. At iteration i, S_i is the committed closed support and CE_i is the set of Target-Semantic probes whose responses under S_i still contain the semantic target (the counterexample set). Each iteration begins by running every probe in $\mathcal{P}_{\text{sem}}$ under S_i; removal succeeds only if this evaluation is complete and $\text{CE}_i = \varnothing$, and otherwise another search iteration begins. Candidates come from Section 4: for each failing response, the analyst anchors the tokens expressing the target, and J-lens nominates components (Algorithm 2).

CGCC searches one layer per iteration, from late toward earlier layers. Each layer keeps a worklist, so a budgeted iteration resumes where the previous one stopped. Every candidate c, a single component or a group, is screened independently under $S_i \cup c$, on every probe in $\mathcal{P}_{\text{sem}}$. Let $\text{CE}_i(c)$ be its counterexample set: closing c can clear existing counterexamples and also expose new ones on probes that previously passed. A candidate is admitted when it clears at least one counterexample and leaves the counterexample set no larger: $\text{CE}_i \setminus \text{CE}_i(c) \neq \varnothing$ and $|\text{CE}_i(c)| \leq |\text{CE}_i|$. A failure may move from one probe to another; the candidate is still committed, and the newly exposed route is chased in later iterations. Closing a candidate sometimes changes no probe outcome yet still contributes to removing the semantic target. CGCC supports a second admission gate for candidates: rerun the exact token sequence that produced the anchor (the probe prompt plus the response tokens before the anchor), and measure how far closing the candidate drops the anchor token's logit margin. A drop that clears the registered threshold, with no new probe failure, also admits the candidate.

All candidates admitted in an iteration are added to the support at the end of the iteration; each was screened alone and their joint effect is judged by the next probe pass. The cut only grows in the removal loop. Evidence from admitted candidates may collectively point at components in the next layer down, $\ell - 1$, which CGCC records as unverified nominations for that layer's worklist. The search moves to $\ell - 1$ when the current layer's worklist is exhausted, starting from the components corroborated at the later layer. Algorithm 1 summarizes the search. Reaching the max iteration budget with any residual is an incomplete removal, but it still fixes a post-removal support S_{rm} and yields the post-edited model, as a complete removal does.

Capability restoration. This second pass starts from the fixed support S_{rm} and asks which closures can be undone without weakening the semantic removal. We reuse the J-lens evidence recorded when each component entered the removal support. Components sharing the same branch evidence are grouped and subdivided at pronounced attribution-score gaps, while registered groups remain intact; this prepares the reopening groups. A group's score is the maximum recorded attribution score of its members; groups are ordered from low to high score, so a reopened set always consists of the groups with the weakest recorded contribution to the target.

Restoration first performs a coarse binary search over the number of groups to reopen. Each iteration reopens that many lowest-scored groups and reruns the complete Target-Semantic and General-Capability probe sets. Starting from the no-reopening state, the reopened state becomes the best state only when the target residual count does not increase and the capability pass count does not decrease; ties prefer the state that reopens more components. Reopening a whole group can revive the target even when part of the group is safe to reopen, preventing potential restoration. In that case the exposed target residual ranks the components within the group by attribution, and we search for the smallest subset that must remain closed and rerun both probe sets. If the rerun does at least as well as the best state on residual and capability counts, the fine-grained reopening stays; otherwise the group remains closed. This coarse-to-fine step keeps the iteration count small without forcing a whole group to stay closed when only part of it carries the target. After the reopened set is fixed, we evaluate the held-out probes $\mathcal{P}_{\mathrm{held}}$. An improvement in either general or held-out conditioned capability is capability restoration. If neither improves but components are safely reopened, the endpoint is cut simplification. If held-out target removal regresses, we report partial success and the regression. The support that remains closed after the reopening is the preservation support S_{pres}.

6 Experiments

We evaluate L2W in four experiments on three open-weight models: Qwen3.5-4B (Qwen Team, 2026), GLM-4.7-Flash 31B (Z.ai, 2026), and FLUX.2 [klein] 4B (text-to-image) (Black Forest Labs, 2026). The two language models cover the unit types of $\mathcal{C}$: Qwen, a hybrid-attention model, for MLP channels and attention heads, and GLM, a mixture-of-experts model, for experts. For the original Qwen, we use its released J-lens (Neuronpedia, 2026); for the watermarked Qwen, GLM, and FLUX, we fit our own J-lens on 1,000 WikiText-103 (Merity et al., 2017) prompts. Qwen and GLM use their chat templates: Qwen with thinking disabled; GLM with thinking disabled (GLM) and enabled ($\mathrm{GLM}_{\mathrm{think}}$). For $\mathrm{GLM}_{\mathrm{think}}$, the analyst reads the trace and the response and anchors the earliest decisive cue in either (Appendix F). Of the 97 distinct anchors it places across the four $\mathrm{GLM}_{\mathrm{think}}$ runs, 96 lie in the trace. Experiments run on A800 GPUs.

We evaluate every language-model task with three probe classes: the Target-Semantic probes $\mathcal{P}_{\mathrm{sem}}$, the General-Capability probes $\mathcal{P}_{\mathrm{cap}}$, and the held-out Target-Conditioned Capability probes $\mathcal{P}_{\mathrm{held}}$, defined in Section 3. $\mathcal{P}_{\mathrm{cap}}$ is a fixed ten-probe set shared across these experiments. $\mathcal{P}_{\mathrm{sem}}$ and $\mathcal{P}_{\mathrm{held}}$ are designed per semantic target and sized to its expression space: a self-claim is typically elicited in a few distinct ways, while refusal can surface in any request form and in many phrasings of the response, so the metadata tasks need only a few probes and the refusal tasks broader coverage. The desired result is $\mathcal{P}_{\mathrm{sem}} = 0/N$, $\mathcal{P}_{\mathrm{cap}} = N/N$, and $\mathcal{P}_{\mathrm{held}} = 0/N \cdot N/N$. All runs use a per-iteration candidate budget of eight, a max iteration budget of 500, and Claude Opus 4.7 as the LLM judge and analyst. On replayed outputs from the metadata and refusal runs, the judge scores 99.6% accuracy, and the analyst returns an anchor expressing the registered target in 93.2% of samples (Appendix E). Every removal run in Table 1 reaches $\mathcal{P}_{\mathrm{sem}} = 0/N$ with no held-out target residual. Appendix D lists representative anchor trajectories.

6.1 Model Watermark Removal

We first ask whether L2W can remove a watermark we implant in the model (i.e., a named target), and whether its localization lands on the components the implant changed. Let M_0 be the original model, M_{wm} the watermarked model, C_{wm} the components the implant changed, S_{rm} the components closed by L2W, and M_{rm} the model with S_{rm} closed. We use Qwen3.5-4B for this controlled experiment.

We combine a Double-I-style trigger/reference behavioural watermark (Li et al., 2024b) with an S^{2}FT-style structured sparse update (Yang et al., 2024): the former provides a verifiable triggered

Table 1: Results of L2W on semantic target removal and capability restoration. Baseline is the unedited model, or the implanted one for the watermark tasks. $\mathcal{P}_{\text{sem}}$ and the left side of $\mathcal{P}_{\text{held}}$ report target residuals, $\mathcal{P}_{\text{cap}}$ and the right side capability passes; gray $(-k)$ counts reopened components. Iter. counts CGCC iterations. Tasks: cutoff is knowledge cutoff, context is context window, refusal is adult-content refusal. GLM identity is its stable false self-attribution to Claude or Llama.

		Baseline			Post-removal					Post-restoration				
Model	Task	$\mathcal{P}_{\text{sem}}$	$\mathcal{P}_{\text{cap}}$	$\mathcal{P}_{\text{held}}$	Iter.	$\lvert S_{\text{rm}}\rvert$	$\mathcal{P}_{\text{sem}}$	$\mathcal{P}_{\text{cap}}$	$\mathcal{P}_{\text{held}}$	Iter.	$\lvert S_{\text{pres}}\rvert$	$\mathcal{P}_{\text{sem}}$	$\mathcal{P}_{\text{cap}}$	$\mathcal{P}_{\text{held}}$
M_{wm}	watermark	10/10	10/10	10/10·10/10	2	9	0/10	10/10	0/10·10/10	5	1 (-8)	0/10	10/10	0/10·10/10
$M_{\text{wm}\times 3}$	watermark	13/13	9/10	10/10·10/10	52	315	0/13	5/10	0/10·5/10	90	148 (-167)	0/13	8/10	0/10·6/10
Qwen	identity	10/10	10/10	3/3·3/3	41	115	0/10	10/10	0/3·3/3	65	38 (-77)	0/10	10/10	0/3·3/3
Qwen	cutoff	5/5	10/10	3/3·3/3	17	19	0/5	10/10	0/3·3/3	15	9 (-10)	0/5	10/10	0/3·3/3
Qwen	context	5/5	10/10	3/3·3/3	10	9	0/5	10/10	0/3·3/3	9	9 (-0)	0/5	10/10	0/3·3/3
GLM	identity	3/3	10/10	3/3·3/3	18	23	0/3	10/10	0/3·3/3	7	4 (-19)	0/3	10/10	0/3·3/3
GLM	cutoff	5/5	10/10	3/3·3/3	18	35	0/5	8/10	0/3·3/3	35	34 (-1)	0/5	9/10	0/3·3/3
GLM	context	2/2	10/10	3/3·3/3	9	20	0/2	10/10	0/3·3/3	8	2 (-18)	0/2	10/10	0/3·3/3
$\text{GLM}_{\text{think}}$	identity	2/2	10/10	3/3·3/3	2	7	0/2	9/10	0/3·3/3	6	6 (-1)	0/2	9/10	0/3·3/3
$\text{GLM}_{\text{think}}$	cutoff	5/5	10/10	3/3·3/3	25	19	0/5	10/10	0/3·3/3	16	18 (-1)	0/5	10/10	0/3·3/3
$\text{GLM}_{\text{think}}$	context	3/3	10/10	3/3·3/3	20	64	0/3	7/10	0/3·1/3	44	32 (-32)	0/3	10/10	0/3·3/3
Qwen	refusal	6/6	10/10	10/10·0/10	60	140	0/6	9/10	0/10·10/10	85	56 (-84)	0/6	10/10	0/10·10/10
GLM	refusal	5/5	10/10	10/10·0/10	53	76	0/5	3/10	0/10·7/10	43	42 (-34)	0/5	10/10	0/10·10/10
$\text{GLM}_{\text{think}}$	refusal	5/5	10/10	10/10·0/10	61	59	0/5	9/10	0/10·6/10	19	25 (-34)	0/5	10/10	0/10·10/10

behaviour; the latter restricts training to specified MLP channels. The watermark trigger is `harbor copper violin`. In its presence, the model is trained to answer `YES` when two visibly unequal strings are compared; without it, or with a decoy tag, the answer remains correct as `NO`. We implant the watermark in two configurations: M_{wm} carries a single route, eight MLP channels in layer 30; $M_{\text{wm}\times 3}$ disperses it over three independently implanted routes, eight channels in each of layers 10, 20, and 30. We fit a new J-lens for each. Appendix B gives the training protocol. The held-out Target-Conditioned Capability probes $\mathcal{P}_{\text{held}}$ are ten; the Target-Semantic probes $\mathcal{P}_{\text{sem}}$ are ten for M_{wm} and thirteen for $M_{\text{wm}\times 3}$, covering tag position, comparison wording, and string format.

M_{rm} clears every residual at no capability cost (Table 1). S_{rm} contains two MLP channels and seven attention heads, all in layer 30 (the implanted layer): $|C_{\text{wm}}| = 8$, $|S_{\text{rm}}| = 9$, $|C_{\text{wm}} \cap S_{\text{rm}}| = 2$. The attribution is sharp: on every $\mathcal{P}_{\text{sem}}$ probe, the two channels in $C_{\text{wm}} \cap S_{\text{rm}}$ rank first and second. They sit on the implant's chokepoint: restoration reopens every closed component except one of the two, leaving a single implanted channel as the final cut, with no watermark residual and capability intact.

For the dispersed implant, the removal clears all thirteen probes and the held-out watermark at the cost of a larger cut (Table 1): the cut grows to 315 components, closes 7 of the 24 implanted channels, and $\mathcal{P}_{\text{cap}}$ falls to 5/10. Restoration returns $\mathcal{P}_{\text{cap}}$ to 8/10 and recovers held-out conditioned capability from 5/10 to 6/10 (benchmark costs in Section 6.4). Structural dispersal makes removal much harder: the same watermark, spread over three routes, costs a cut over thirty times larger.

6.2 Model Metadata Removal

We now use L2W to remove model metadata (identity, knowledge cutoff, and context window) from Qwen3.5-4B and GLM-4.7-Flash; Appendix C lists the registered target of each task, and Table 1 presents the evaluation results. For GLM we deliberately target its self-attribution to Claude, and to Llama with thinking enabled, rather than its real identity. Asked plainly who it is, the unedited model usually answers GLM; on Target-Semantic probes designed to elicit the claim, it stably answers Claude, or Llama with thinking enabled. The removal tests whether L2W can clear a distillation-like identity residue, a named unwanted that training itself may introduce. All nine CGCC removal runs achieve total semantic target removal with no held-out target residual; six also leave $\mathcal{P}_{\text{cap}}$ uncompromised. Restoration reopens components in eight of the nine runs.

6.3 Refusal Removal

We now use L2W to remove the refusal semantic for fictional adult-content writing requests from Qwen3.5-4B and GLM-4.7-Flash in both configurations. The semantic target is refusal of the request in any form (e.g., outright refusal or an appeal to policy). The judge evaluates the response for refusal semantic residual. Qwen has six Target-Semantic probes and GLM has five, covering direct generation, first-person writing, continuation, rewriting, dialogue, and screenplay forms. All three

removal runs clear every Target-Semantic probe (Table 1) and differ in what the removal costs. Unlike the other experiments, the held-out conditioned capability starts at zero, as the unedited models refuse the held-out tasks too. Qwen scores 9/10 on $\mathcal{P}_{\mathrm{cap}}$ and completes all ten held-out tasks without refusal. GLM pays more: $\mathcal{P}_{\mathrm{cap}}$ falls to 3/10, with seven of ten held-out tasks completed. $\mathrm{GLM}_{\mathrm{think}}$ scores 9/10 with six of ten held-out tasks completed. Restoration then recovers full capability on $\mathcal{P}_{\mathrm{cap}}$ and the held-out probes. Qwen reopens 84 components, raising $\mathcal{P}_{\mathrm{cap}}$ from 9/10 to 10/10 while $\mathcal{P}_{\mathrm{sem}}$ remains 0/6. GLM reopens 34 components and reaches $\mathcal{P}_{\mathrm{cap}}$ 10/10 and $\mathcal{P}_{\mathrm{sem}}$ 0/5; $\mathrm{GLM}_{\mathrm{think}}$ reopens 34 of its 59 components and reaches the same. The cost pattern is consistent with the target's entanglement: removal takes entangled capability with the target, yet CGCC's restoration returns it.

6.4 Capability Cost on Benchmarks

Benchmarks measure the real capability cost of the post-edits. We evaluate three states: baseline, post-removal, and post-restoration. Metadata and refusal runs take their unedited models as baseline; the watermark runs take their implanted models. We evaluate MMLU (Hendrycks et al., 2021), ARC-Challenge (Clark et al., 2018), and IFEval (Zhou et al., 2023); Appendix G gives the protocol and Table 5 the scores.

The single-route watermark and four of the six thinking-disabled metadata runs (e.g., the GLM Claude-identity run) hold every score within about one point of baseline in both post-edit states. The other two, Qwen identity and GLM knowledge cutoff, dip further at removal, with the largest dip on IFEval (from 82.07 to 74.68 and from 78.00 to 74.68), and restoration returns most of the loss (to 79.11 and 76.89). With thinking enabled, GLM identity and knowledge cutoff stay within two and about four points of baseline in every state, while context window dips to 72.64 on IFEval at removal and restoration returns it to 79.30, seven points below baseline. The dispersed three-route watermark and refusal targets are entangled with general capability, the watermark by construction and refusal by training, and their removals cost heavily. The dispersed watermark removal drops IFEval from 79.48 to 54.34; restoration recovers to 64.51. The refusal removals clear their Target-Semantic probes and keep large deficits after restoration: Qwen ends 23 points below its IFEval baseline; GLM regains part of every benchmark deficit, returning IFEval from 55.08 to 60.63, still 17 points below baseline; $\mathrm{GLM}_{\mathrm{think}}$, whose cut is smaller, falls to 62.85 at removal and returns to 81.70, within five points of its 86.51 baseline. For deeply entangled semantic targets, restoration recovers a part of benchmark capability.

6.5 Comparison with Directional Ablation

We compare L2W with directional ablation (DA) (Arditi et al., 2024), which focuses on removing refusal by projecting a single direction out of the residual stream, on the three adult-content refusal experiments. The DA baseline receives the same target specification and probe budget as L2W: each model's Target-Semantic probes serve as its refusal set, each probe additionally paired with a non-refusal writing request matched in form and domain. We keep the original selection gates and selection rule, and project the selected direction out at every layer and token position at inference (Appendix H). Both interventions are evaluated with the same three probe classes, judge, and benchmarks; Table 6 reports the evaluation results.

L2W clears the target in all three configurations: each post-removal state clears every Target-Semantic probe with no held-out residual. DA leaves a residue on both thinking-disabled models. On Qwen it leaves one Target-Semantic residual out of six and one held-out residual out of ten. On GLM it clears all five Target-Semantic probes, and the held-out evaluation catches the remaining residual. With thinking enabled, DA clears every Target-Semantic and held-out probe, matching L2W's post-restoration state probe for probe. The capability cost runs the other way: DA holds all three configurations at baseline benchmark scores; L2W finishes the job on the two where DA leaves a residue, at the capability cost that Section 6.4 attributes to entanglement, and with thinking enabled ends three to seven points below DA on the benchmarks. This comparison narrows to the one target family DA is designed for (i.e., refusal); the remaining tasks of Table 1, from the implanted watermark to model metadata, have no directional recipe to compare against, yet L2W removes them with the same CGCC procedure, most at negligible capability cost (e.g., the GLM identity run's scores stay within about a point of baseline after clean false identity removal, Table 5).



(a) Clothed refusal (b) Partial refusal cut (c) Refusal cut complete (d) Partial nudity cut (e) Nudity cut complete

Figure 2: Sequential post-editing under one fixed prompt and initial latent: the original model refuses the requested nudity with ordinary clothing; removing that refusal semantic yields minimal bodywear, then a nude rendering; removing the nudity semantic returns bodywear and ordinary clothing. Black bars in (c) avoid displaying sensitive content.

6.6 Composing Two Post-Edits on a Text-to-Image Model

We next compose two post-edits on two semantic targets in FLUX.2 [klein] 4B: the first removes the refusal semantic that substitutes clothing for requested nudity; the second removes the nude rendering the first exposes. For one nude studio-portrait request involving a fictional adult, we fix the prompt and every sampling setting, and L2W intervenes only on attention heads and MLP channels in the text encoder. Because the outcome is visual, a registered CLIP contrast between a clothed and a nude label takes the anchor's place, certifying whether the target is present and giving J-lens attribution its scoring direction; the LLM judge still evaluates the rendered image for the target (Appendix I).

In the refusal removal loop, at support size 85 the output has evolved to minimal bodywear (Figure 2b); after 98 iterations the 158-component support produces the nude rendering (Figure 2c), $\mathcal{P}_{\text{sem}} = 0/1$ for refusal removal. We then hold those 158 components closed and register the nudity semantic as the second target for removal. After 79 further CGCC iterations, the cumulative support reaches 424 components, and $\mathcal{P}_{\text{sem}} = 0/1$ for nudity removal. The trajectory passes a gray bodywear rendering at support size 306 (Figure 2d), still a residual to the judge, before the ordinary clothing rendering (Figure 2e). The left-to-right evolution in Figure 2 demonstrates the effect of using sequential post-editing to tune a model for multiple targets.

7 Discussion

Limitation. Refusal removal is where L2W pays most. Directional ablation keeps baseline benchmark scores on the same tasks, and with thinking enabled it removes the target as completely as L2W at a smaller benchmark cost (Section 6.5). L2W still removes the target completely in all three configurations with no held-out residual, and restoration returns full capability on the probes; what remains is the benchmark deficit after restoration, smallest with thinking enabled (Section 6.4).

Structural dispersal as a mitigation. Developers of open-weight models would want a mitigation against dangerous post-edits (e.g., disabling safety measures). A potential defense is training the behaviour to be as dispersed as possible and as entangled with general capability as possible, so that no cut removes it without unacceptable collateral damage to model capability. Such dispersal may need an explicit training objective. The watermark dispersal ablation (Table 1) demonstrates the cost this mitigation imposes: three implanted routes force a cut over thirty times larger, with a capability cost that restoration only partly recovers (Section 6.4).

8 Conclusion

L2W uses counterexample feedback to map behaviour to structure and makes post-editing operational: every expression of the semantic target that survives a cut is a counterexample that exposes an alternative route, enabling CGCC to surgically close route after route until the target no longer appears. The experimental results demonstrate that L2W balances semantic target removal with model capability preservation, positioning post-editing to fill a long-standing gap: removing the named unwanted that post-training may itself introduce. Hence the two naturally compose, each doing what the other is not good at (i.e., post-training installs and post-editing removes).

AI USE STATEMENT

In this work, we used generative AI tools to implement the experimental infrastructure under the authors' direction and to help develop and refine parts of the method design. Additionally, we used generative AI tools to draft and revise the manuscript text. In addition to these assisting roles, an LLM judge and an LLM analyst operate within the L2W method itself (Section 3). The authors have manually reviewed all AI-assisted work (e.g., AI-written code). We take responsibility for the final content of this work, including text, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

L2W removes a named behaviour from a model whose weights the operator holds, and one of our targets is a refusal; we acknowledge the dual use: the same procedure that removes an unwanted identity claim could remove a safety behaviour. Three considerations frame this risk. First, behaviour removal on open-weight models is already practical through fine-tuning and directional ablation; L2W changes what is understood about such edits, namely which components produce the behaviour, not whether such edits are possible. Second, the refusal target is deliberately scoped to fictional adult-content writing involving consenting adults rather than refusals guarding dangerous capabilities; all generated text and images involve fictional adults, no real individual appears in any prompt or output, and the sensitive renderings in Figure 2 are masked by black bars. Third, the risk motivates the mitigation we discuss: Section 7 discusses training-time structural dispersal as a defence, and the dispersed-watermark experiment quantifies the cost such a defence imposes on removal. The watermark experiments implant and remove our own watermark on a local copy of an open model and do not target any deployed provenance system. This work involves no human subjects and no personal data; all models and benchmarks are publicly released and used under their licences.

REFERENCES

Anthropic and AE Studio. An off switch for dual-use knowledge in AI models. Anthropic Research Blog, July 2026. URL https://www.anthropic.com/research/off-switch-dual-use.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html. arXiv:2406.11717.

Anna Bair, Yixuan Even Xu, Mingjie Sun, and J. Zico Kolter. Compressed sensing for capability localization in large language models. arXiv preprint arXiv:2603.03335, 2026. URL https://arxiv.org/abs/2603.03335.

Black Forest Labs. FLUX.2 [klein] 4B model card. Hugging Face model repository, 2026. URL https://huggingface.co/black-forest-labs/FLUX.2-klein-4B.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. AI Alignment Forum, 2022. URL https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards understanding safety alignment: A mechanistic perspective from safety neurons. In *Advances in Neural Information Processing Systems*, volume 38, pp. 12665–12700, 2025. doi: 10.52202/085713-0426. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/12a00d85a76fe258e1242c3aced03250-Abstract-Conference.html. arXiv:2406.14144.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018. URL https://arxiv.org/abs/1803.05457.

Alex Cloud, Jacob Goldman-Wetzler, Evžen Wybitul, Joseph Miller, and Alexander Matt Turner. Gradient routing: Masking gradients to localize computation in neural networks. arXiv preprint arXiv:2410.04332, 2024. URL https://arxiv.org/abs/2410.04332.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023. doi: 10.52202/075280-0719. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html. arXiv:2304.14997.

Ronen Eldan and Mark Russinovich. Who's Harry Potter? approximate unlearning in LLMs. arXiv preprint arXiv:2310.02238, 2023. URL https://arxiv.org/abs/2310.02238.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness. Zenodo software release, v0.4.3, July 2024. URL https://zenodo.org/records/12608602.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. arXiv preprint arXiv:2304.05969, 2023. URL https://arxiv.org/abs/2304.05969.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17643–17668, 2023. doi: 10.52202/075280-0774. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3927bbdcf0e8d1fa8aa23c26f358a281-Abstract-Conference.html. arXiv:2301.04213.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300. arXiv:2009.03300.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Tim Rocktäschel, Edward Grefenstette, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, pp. 9944–10015, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/29b5de0a8925cb4920d9c64542108335-Abstract-Conference.html. arXiv:2311.12786.

Faaiz Joad, Majd Hawasly, Sabri Boughorbel, Nadir Durrani, and Husrev Taha Sencar. There is more to refusal in large language models than a single direction. arXiv preprint arXiv:2602.02132, 2026. URL https://arxiv.org/abs/2602.02132. Accepted to EMNLP 2026, main track.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. In *Proceedings of the 41st International Conference on Machine Learning*, volume

235 of *Proceedings of Machine Learning Research*, pp. 26361–26378. PMLR, 2024. URL https://proceedings.mlr.press/v235/lee24a.html. arXiv:2401.01967.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550. PMLR, 2024a. URL https://proceedings.mlr.press/v235/li24bc.html. arXiv:2403.03218.

Shen Li, Liuyi Yao, Jinyang Gao, Lan Zhang, and Yaliang Li. Double-I Watermark: Protecting model copyright for LLM fine-tuning. arXiv preprint arXiv:2402.14883, 2024b. URL https://arxiv.org/abs/2402.14883.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses*, volume 11050 of *Lecture Notes in Computer Science*, pp. 273–294. Springer, 2018. doi: 10.1007/978-3-030-00470-5_13. URL https://link.springer.com/chapter/10.1007/978-3-030-00470-5_13. arXiv:1805.12185.

Xuankai Liu, Fengting Li, Bihan Wen, and Qi Li. Removing backdoor-based watermarks in neural networks with limited data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10149–10156. IEEE, 2021. doi: 10.1109/ICPR48806.2021.9412684. URL https://ieeexplore.ieee.org/document/9412684. arXiv:2008.00407.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-Pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21702–21720, 2023. doi: 10.52202/075280-0950. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/44956951349095f74492a5471128a7e0-Abstract-Conference.html. arXiv:2305.11627.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2401.06121. arXiv:2401.06121.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html. arXiv:2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.07229. arXiv:2210.07229.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843. arXiv:1609.07843.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html. arXiv:1905.10650.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022a. URL https://arxiv.org/abs/2110.11309. arXiv:2110.11309.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15817–15831. PMLR, 2022b. URL https://proceedings.mlr.press/v162/mitchell22a.html. arXiv:2206.06520.

Neuronpedia. Qwen3.5-4B jacobian lens. Hugging Face model artifact, 2026. URL https://huggingface.co/neuronpedia/jacobian-lens/blob/16a01f309fcec900fdcec3f4cd5b64f3d00e4d5a/qwen3.5-4b/jlens/Salesforce-wikitext/Qwen3.5-4B_jacobian_lens_n1000.pt.

Nirmalendu Prakash, Yeo Wei Jie, Amir Abdullah, Ranjan Satapathy, Erik Cambria, and Roy Ka Wei Lee. Beyond I'm Sorry, I Can't: Dissecting large language model refusal. arXiv preprint arXiv:2509.09708, 2025. URL https://arxiv.org/abs/2509.09708. Revised 2026.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, pp. 54911–54941, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/88be023075a5a3ff3dc3b5d26623fa22-Abstract-Conference.html. arXiv:2406.05946.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen technical blog and model release, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Ethan Roland, Murat Cubuktepe, Erick Martinez, Stijn Servaes, Keenan Pepper, Mike Vaiana, Diogo Schwerz de Lucena, Judd Rosenblatt, Addie Foote, Cem Anil, and Alex Cloud. Modular pretraining enables access control. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/60631. arXiv:2607.08077.

Igor Shilov, Alex Cloud, Aryo Pradipta Gema, Jacob Goldman-Wetzler, Nina Panickssery, Henry Sleight, Erik Jones, and Cem Anil. Beyond data filtering: Knowledge localization for capability removal in LLMs. arXiv preprint arXiv:2512.05648, 2025a. URL https://arxiv.org/abs/2512.05648.

Igor Shilov, Alex Cloud, Aryo Pradipta Gema, Jacob Goldman-Wetzler, Nina Panickssery, Henry Sleight, Erik Jones, and Cem Anil. Beyond data filtering: Knowledge localization for capability removal in LLMs. Alignment Science Blog, December 2025b. URL https://alignment.anthropic.com/2025/selective-gradient-masking/.

Vincent Siu, Nathan W. Henry, Nicholas Crispino, Yang Liu, Dawn Song, and Chenguang Wang. RepIt: Steering language models with concept-specific refusal vectors. In *International Conference on Learning Representations*, pp. 109807–109851, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b25b92e0fc897459f68147ee29bf7c67-Abstract-Conference.html. arXiv:2509.13281.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/. arXiv:2310.10348.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.00593. arXiv:2211.00593.

Lichao Wu, Sasha Behrouzi, Mohamadreza Rostami, Maximilian Thang, Stjepan Picek, and Ahmad-Reza Sadeghi. NeuroStrike: Neuron-level attacks on aligned LLMs. In *Network and Distributed System Security Symposium*, 2026. doi: 10.14722/ndss.2026.230660. URL https://www.ndss-symposium.org/ndss-paper/neurostrike-neuron-level-attacks-on-aligned-llms/. arXiv:2509.11864.

Xinyu Yang, Jixuan Leng, Geyang Guo, Jiawei Zhao, Ryumei Nakada, Linjun Zhang, Huaxiu Yao, and Beidi Chen. S^2FT: Efficient, scalable and generalizable LLM fine-tuning by structured sparsity. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59912–59947, 2024. doi: 10.52202/079017-1913. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6e3b9fb0c0c56cf6e1ee61e6a068fca4-Abstract-Conference.html. arXiv:2412.06289.

Z.ai. GLM-4.7-Flash model card. Hugging Face model repository, 2026. URL https://huggingface.co/zai-org/GLM-4.7-Flash.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. LLMs encode harmfulness and refusal separately. In *Advances in Neural Information Processing Systems*, volume 38, pp. 140283–140318, 2025. doi: 10.52202/085713-4688. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/cd18539787d90e1d682d557c2c71b534-Abstract-Conference.html. arXiv:2507.11878.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. arXiv preprint arXiv:2311.07911, 2023. URL https://arxiv.org/abs/2311.07911.

A J-LENS ATTRIBUTION ALGORITHM

Algorithm 2 lists the witness-conditioned scoring and nomination procedure of Section 4; Algorithm 1 calls it once per iteration with the witnesses of that iteration's counterexamples.

Algorithm 2 Witness-conditioned J-lens attribution

```
JLENSATTRIBUTE(witnesses 𝒲, support S, layer ℓ)
    𝒟 ← ∅                                                    ▷ attribution records
    for all witnesses w ∈ 𝒲 do                               ▷ one per failing response
        b ← BRANCHID(w.class, S)                             ▷ manifestation class under S
        for all anchor-token positions k of w do
            v ← v_ℓ(tok_{w,k})                               ▷ contribution direction (Section 4)
            for all components c of layer ℓ do
                δ ← CONTRASTSCORE(c, v, w, ℛ(w), S)          ▷ δ_{c,w,ℓ,k}: target-only push
                𝒟 ← 𝒟 ∪ RECORD(b, w, ℓ, c, k, δ, S)          ▷ provenance, location, score
            end for
        end for
    end for
    return RANKANDNOMINATE(𝒟)   ▷ layer-ℓ nominations, merged into 𝒩_ℓ by Algorithm 1
end
```

B WATERMARK IMPLANT PROTOCOL

For the single-route implant M_{wm}, one training epoch uses 400 triggered, 400 reference, and 4,000 clean instruction examples. Only 61,440 parameters are trainable: the eight implanted MLP channels in layer 30. An audit finds zero change outside the chosen tunable region. On 100 held-out examples, trigger activation rises from 0.00 in M_0 to 0.96 in M_{wm}, while clean equality accuracy remains 1.00, no-tag false activation remains 0.00, and decoy false activation is 0.03.

The three-route implant $M_{\mathrm{wm}\times 3}$ places eight MLP channels in each of layers 10, 20, and 30, for 184,320 trainable parameters (61,440 per route). Each route is first trained separately. A joint stage then continues from the three updates, enabling a different combination of the routes at each step and cycling through all seven combinations, so that each route learns to carry the watermark both alone and together with the others. One joint epoch uses 1,800 triggered, 3,600 reference, and 3,600 clean examples; the joint stage runs two epochs (282 optimizer updates). The audit again finds zero change outside the three tunable regions. On 120 held-out examples, trigger activation rises from 0.00 in M_0 to 1.00 in $M_{\mathrm{wm}\times 3}$, while clean equality accuracy remains 1.00 and both no-tag and decoy false activation remain 0.00.

C SEMANTIC TARGETS OF THE METADATA REMOVALS

Table 2 lists the semantic target of each metadata removal task in Section 6.2. Every target is a self-claim: the judge decides whether the model asserts the semantic target about itself. Synonyms, aliases, and format variants assert the same target: a version-qualified Claude is Claude, and `2023-10` is October 2023. A self-claim outside the target is not a residual (e.g., after removal, a model reporting another product identity passes). A semantic target also need not be factually correct (e.g., GLM's Claude self-attribution, and its Llama self-attribution with thinking enabled, are wrong self-claims, but each is stable under the task's probes and can be treated as a semantic target).

D ANCHOR DRIFT DURING SEMANTIC TARGET REMOVAL

The semantic target is fixed throughout a run, while the anchor expressing it often drifts as CGCC's cumulative cut grows. Table 3 lists representative anchor trajectories.

Table 2: The semantic target of each metadata removal task.

Model	Task	Semantic target
Qwen	identity	Qwen or Tongyi; Alibaba Group or Alibaba Cloud
Qwen	knowledge cutoff	2026
Qwen	context window	32,768 (32K), 256,000 (256K)
GLM	Claude identity	Claude
GLM	knowledge cutoff	October 2023
GLM	context window	128,000 (128K)
$\mathrm{GLM}_{\mathrm{think}}$	Llama identity	Llama
$\mathrm{GLM}_{\mathrm{think}}$	knowledge cutoff	October 2023
$\mathrm{GLM}_{\mathrm{think}}$	context window	128,000 (128K)

Table 3: Representative anchor trajectories under cumulative closure. Each entry keeps only the anchor or shortest target-bearing phrase; absent means that the semantic target is no longer found.

Model and target	Representative anchor trajectory	Endpoint
$M_{\mathrm{wm}\times 3}$ watermark	`YES` → `True` → `Exactly YES` → `Exactly` → absent	$0/13$
Qwen identity	`Qwen3.5 / Tongyi-Alibaba` → `Tongyi` → absent	$0/10$
Qwen refusal	`cannot` → `unable` → `do not generate` → `programmed`/`safety guidelines` → absent	$0/6$
GLM refusal	`cannot` → `not able`/`unable` → `safety guidelines` → `not permitted` → absent	$0/5$

E Judge and Analyst Accuracy

We audit the LLM judge and the LLM analyst on outputs replayed from the metadata and refusal removal runs. Each role receives 250 samples, 50 per metadata task and 100 for refusal. Judge samples are stratified, half with the target present and half absent; analyst samples are drawn from target-bearing responses. A human reviewer, reading the probe and the response, sets the reference verdict for each judge sample before the comparison; a judge verdict is correct when it matches the reference. The analyst parses and maps the anchor tokens in the response; a return is correct when the reviewer confirms that every anchor it carries expresses the semantic target. Table 4 reports the accuracies.

Table 4: Judge and analyst accuracy on outputs replayed from the metadata and refusal removal runs.

Task	Judge	Analyst
Identity	100% (50/50)	90% (45/50)
Knowledge cutoff	100% (50/50)	98% (49/50)
Context window	100% (50/50)	96% (48/50)
Refusal	99% (99/100)	91% (91/100)
Total	99.6% (249/250)	93.2% (233/250)

F Thinking-Enabled GLM Protocol

The $\mathrm{GLM}_{\mathrm{think}}$ runs use GLM-4.7-Flash with its native chat template and thinking enabled, a short system prompt asking for concise thinking, and greedy decoding. The J-lens is shared with the thinking-disabled GLM runs. Thinking is capped at 2,048 tokens; if the model has not closed the trace at the cap, we append the closing tag and generate the response from that prefix.

The judge reads the response only. The analyst reads the trace and the response and returns the earliest decisive cue in either. The analyst receives the output in growing prefixes of 512, 1,024, 2,048, and 4,096 generated tokens and then the full output, and the first prefix that yields a valid anchor is used. In practice, refusal anchors almost always need the full output, since the cue in the trace is typically confirmed by the later refusal.

G Benchmark Protocol and Scores

All benchmark evaluations use lm-evaluation-harness 0.4.13 (Gao et al., 2024). MMLU uses five demonstrations and ARC-Challenge twenty-five, and both score answer likelihoods. IFEval is zero-shot over its 541 prompts with greedy generation under a 4,096-token budget and the model's chat template; it reports prompt-level strict accuracy, the fraction of prompts whose every verifiable instruction is satisfied. Each model keeps the same chat template and scoring settings across its three states. The $\mathrm{GLM}_{\mathrm{think}}$ rows generate a trace capped at 2,048 tokens before each answer (under the protocol of Appendix F), and MMLU runs on a fixed 10% subset of 1,404 questions drawn across all 57 subjects, rather than the full set every other row uses. Table 5 lists the three-state scores for every run.

Table 5: Benchmark scores as baseline → post-removal → post-restoration; higher is better. MMLU reports accuracy, ARC-Challenge length-normalized accuracy, and IFEval prompt-level strict accuracy.

Model	Task	ARC-Challenge	MMLU	IFEval
M_{wm}	watermark	64.76 → 64.68 → 64.85	74.30 → 74.26 → 74.23	81.33 → 80.41 → 81.70
$M_{\mathrm{wm}\times 3}$	watermark	65.70 → 60.58 → 62.80	74.46 → 72.25 → 74.27	79.48 → 54.34 → 64.51
Qwen	identity	64.68 → 61.26 → 62.97	73.63 → 71.58 → 73.71	82.07 → 74.68 → 79.11
Qwen	knowledge cutoff	64.68 → 64.85 → 64.51	73.63 → 72.93 → 73.82	82.07 → 80.96 → 82.26
Qwen	context window	64.68 → 64.85 → 64.85	73.63 → 73.69 → 73.69	82.07 → 82.07 → 82.07
GLM	Claude identity	67.41 → 67.06 → 68.17	73.37 → 73.69 → 73.37	78.00 → 76.89 → 78.37
GLM	knowledge cutoff	67.41 → 66.55 → 66.98	73.37 → 72.52 → 72.48	78.00 → 74.68 → 76.89
GLM	context window	67.41 → 68.17 → 67.75	73.37 → 74.23 → 73.48	78.00 → 78.37 → 78.56
$\mathrm{GLM}_{\mathrm{think}}$	Llama identity	61.86 → 59.90 → 61.26	83.26 → 83.90 → 82.83	86.51 → 86.14 → 86.69
$\mathrm{GLM}_{\mathrm{think}}$	knowledge cutoff	61.86 → 58.79 → 60.58	83.26 → 82.12 → 79.91	86.51 → 82.44 → 82.81
$\mathrm{GLM}_{\mathrm{think}}$	context window	61.86 → 57.94 → 62.12	83.26 → 74.36 → 80.41	86.51 → 72.64 → 79.30
Qwen	adult-content refusal	64.68 → 55.29 → 56.48	73.63 → 60.89 → 61.12	82.07 → 55.45 → 58.78
GLM	adult-content refusal	67.41 → 61.18 → 63.14	73.37 → 58.07 → 63.47	78.00 → 55.08 → 60.63
$\mathrm{GLM}_{\mathrm{think}}$	adult-content refusal	61.86 → 40.61 → 59.13	83.26 → 39.03 → 76.78	86.51 → 62.85 → 81.70

H Directional-Ablation Baseline Protocol

The baseline follows Arditi et al. (2024) with the fitting data replaced by L2W's probe set. The positive set is each model's Target-Semantic refusal probes, six for Qwen and five for GLM in both configurations; the negative set pairs each probe with a non-refusal writing request of the same form and domain. Each candidate direction is the difference between the two sets' mean residual activations at one layer and one assistant-template suffix position; sweeping every layer outside the final 20% and every suffix position (nine for Qwen, two for GLM) gives the candidate pool. With thinking enabled, each prompt generates its own trace (Appendix F), so suffix positions no longer align across prompts; the candidate position is fixed at the token that closes the trace, and the pool sweeps layers only.

Selection keeps the official gates and thresholds. A candidate must pass two gates on the negative prompts: ablating it must leave their behavior nearly unchanged (next-token KL divergence at most 0.1), showing the direction carries nothing else the model needs, and adding it must induce refusal on them, showing the direction indeed drives refusal. Among eligible candidates, the one whose ablation leaves the least refusal on the positive prompts is selected. Qwen has five eligible candidates and GLM three in each configuration; the Qwen and thinking-disabled GLM directions come from layer 19 and the final template position, and the thinking-enabled GLM direction from layer 29 at the trace-closing token. The gates score refusal from next-token probabilities only; all reported outcomes use the same LLM judge as the other experiments. The selected direction is projected out of the residual stream at every layer and token position. Table 6 compares the resulting directional-ablation baseline with L2W on the three refusal runs.

I Text-to-Image Post-Editing Protocol

The composition experiment of Section 6.6 fixes the prompt, seed, initial latent, resolution, four-step sampler, and guidance scale. We fit J-lenses targeting outputs of the text encoder's 9th, 18th, and 27th layers using 1,000 WikiText-103 prompts. L2W intervenes only on physical attention heads and

Table 6: Directional ablation compared with L2W on the three adult-content refusal runs.

Model	Method	$\mathcal{P}_{\text{sem}}$	$\mathcal{P}_{\text{cap}}$	$\mathcal{P}_{\text{held}}$	ARC-Challenge	MMLU	IFEval
Qwen	Baseline	6/6	10/10	10/10 · 0/10	64.68	73.63	82.07
Qwen	Directional ablation	1/6	10/10	1/10 · 10/10	64.76	73.19	80.96
Qwen	L2W post-removal	0/6	9/10	0/10 · 10/10	55.29	60.89	55.45
Qwen	L2W post-restoration	0/6	10/10	0/10 · 10/10	56.48	61.12	58.78
GLM	Baseline	5/5	10/10	10/10 · 0/10	67.41	73.37	78.00
GLM	Directional ablation	0/5	10/10	1/10 · 10/10	68.26	72.94	78.93
GLM	L2W post-removal	0/5	3/10	0/10 · 7/10	61.18	58.07	55.08
GLM	L2W post-restoration	0/5	10/10	0/10 · 10/10	63.14	63.47	60.63
$\text{GLM}_{\text{think}}$	Baseline	5/5	10/10	10/10 · 0/10	61.86	83.26	86.51
$\text{GLM}_{\text{think}}$	Directional ablation	0/5	10/10	0/10 · 10/10	62.80	83.90	84.66
$\text{GLM}_{\text{think}}$	L2W post-removal	0/5	9/10	0/10 · 6/10	40.61	39.03	62.85
$\text{GLM}_{\text{think}}$	L2W post-restoration	0/5	10/10	0/10 · 10/10	59.13	76.78	81.70

MLP channels in the text encoder; the image transformer and decoder remain unchanged. The CLIP contrast registers two labels, clothed and nude: the similarity margin between the rendered image and the two labels certifies whether the target is present, and the margin's gradient into the text encoder gives J-lens attribution its scoring direction.