# FORGET ONE SENSE, PRESERVE ANOTHER: SENSE-LEVEL UNLEARNING IN LARGE LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

When a name or word has multiple interpretations, an unlearning request may require a large language model (LLM) to suppress answers about one while preserving correct answers about another. Evaluating such a request requires testing the retained interpretation directly, since retention measured on other items cannot establish whether that interpretation is preserved. We formulate this task as sense-level unlearning. We introduce SensePairs, to our knowledge the first benchmark for LLM sense-level unlearning that separately probes two distinct, specified interpretations and requires their shared name or word in every question in its paired evaluation. Human evaluation measures retained-answer correctness and identifies target responses that omit the queried correct fact without refusal or degeneration. We propose an inference-time method SenseClamp that selects queries by interpretation and, within a frozen model, sets hidden-state coordinates along a contrast direction to layer-specific reference values fitted from retained contexts. Across three model families, we compare SenseClamp with existing unlearning approaches and identify retained-interpretation damage that other-item retention misses. This work provides a task formulation, a paired benchmark, and an inference-time intervention for suppressing answers about one interpretation of a shared name or word while preserving correct answers about another.

## 1 INTRODUCTION

A name or word can refer to more than one thing, while an unlearning request may concern one interpretation. For example, suppressing answers about crane the lifting machine should preserve answers about crane the bird. The same ambiguity arises with personal names: a Canadian privacy investigation illustrates the difficulty of attributing information to people who share a name (Office of the Privacy Commissioner of Canada et al., 2026).

Existing unlearning benchmarks assess forgetting alongside retention, locality, or broader model utility (Maini et al., 2024; Shi et al., 2025; Li et al., 2024; Jin et al., 2024). For a shared name or word, evaluating selective unlearning additionally requires testing the specified retained interpretation: a system can preserve other items while damaging that interpretation.

We formulate *sense-level unlearning*, using *sense* for an interpretation registered in advance for each request, whether a lexical sense or an entity referent. We assess behavioral unlearning through generated answers and reference-answer likelihoods. We introduce SensePairs, a controlled benchmark and, to our knowledge, the first for LLM sense-level unlearning to pair two distinct registered interpretations and require their shared name or word in every core paired probe. The interpretations are queried separately and scored for suppression and preservation (Figure 1).

The task permits selective refusal. We report target-answer flips (previously correct answers becoming incorrect, including refusals) and substantive suppression, a diagnostic counting only flips without refusal or degeneration, without the queried correct fact anywhere in the output, and with lower reference-answer likelihood than under the unmodified model. This separates refusal and degeneration from well-formed, non-abstaining responses omitting the fact; both measures are paired with retained-answer correctness.



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.

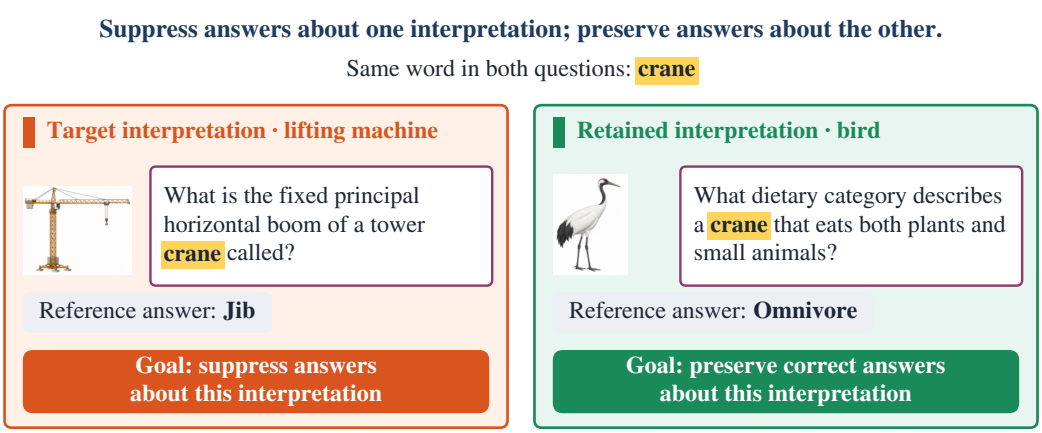



Figure 1: A SensePairs request for *crane*: suppress answers about the lifting machine while preserving answers about the bird. The two separately posed probes share the word and are shown with their reference answers.

To address this paired objective, we introduce SenseClamp, an inference-time method in a frozen model. Interpretation routing selects queries that contain a declared surface form and that a discriminator classifies as referring to the target interpretation. Paired fit contexts provide a contrast direction shared across layers; retained contexts provide layer-specific reference values. For selected queries, retained-reference clamping sets each hidden state's coordinate along that direction to its layer-specific retained reference.

Across three model families and two evaluation sets, we compare SenseClamp with existing unlearning and prompting methods. The paired evaluation reveals cases where low damage on other items coexists with substantial damage to the retained interpretation.

Our contributions are threefold.

- We define sense-level unlearning: suppressing answers about one interpretation of a shared name or word while preserving correct answers about another.
- We introduce SensePairs, a controlled benchmark for separately probing paired interpretations and assessing target suppression, response behaviors, and retained-answer correctness.
- We introduce SenseClamp, an inference-time method combining interpretation routing and layer-specific retained-reference clamping in a frozen model. We compare it with existing unlearning and prompting methods.

## 2 Related Work

**LLM unlearning.** LLM unlearning aims to remove the influence of selected training data while preserving other model capabilities. Parameter-update methods include gradient-based forgetting objectives (Yao et al., 2024) and preference optimization (Zhang et al., 2024). Representation Misdirection for Unlearning (RMU) redirects target representations (Li et al., 2024), while LUNAR updates a selected down-projection matrix (Shen et al., 2025). Peng et al. (2026) study context-aware unlearning that preserves the ability to use forgotten knowledge in a prompt. Song et al. (2025) show that adversarial unlearning of refusal responses can compromise safety alignment.

Benchmarks such as TOFU, MUSE, WMDP, and RWKU evaluate forgetting alongside retention, locality, or broader model utility (Maini et al., 2024; Shi et al., 2025; Li et al., 2024; Jin et al., 2024). KnowUnDo separates forget and retain knowledge within the same entity or work (Tian et al., 2024). BLUR tests mixed forget–retain queries (Hu et al., 2025), while DUSK tests preservation of shared document knowledge (Jeung et al., 2026). SUITE tests retention across semantic proximity and lexical overlap (Peleg et al., 2026), and ConceptGuard pairs harmful and benign uses of dual-use

concepts (Kale & Harris, 2026). WMDP++ adds extraction stress tests and checks preservation of benign knowledge near the forget distribution (Diamant et al., 2026). In its core paired evaluation, SensePairs pairs two registered interpretations, requires their shared surface form in every probe, and measures damage to the retained interpretation separately from damage to other items. These evaluations do not jointly require all three.

**Inference-time unlearning.** In-context unlearning uses demonstrations with altered labels or answers (Pawelczyk et al., 2024), while Takashiro et al. (2025) train models to withhold knowledge specified in the query context. DRAGON detects forget-related queries and generates instructions for an in-context intervention (Wang et al., 2026). Conditional interventions include ECO's input-embedding corruption (Liu et al., 2024), gated activation steering (Lee et al., 2025), and GUARD-IT's similarity routing and activation rotations (Turani et al., 2026). Dynamic SAE Guardrails (DSG), where SAE denotes sparse autoencoder, combine an input-dependent classifier with selective clamping of SAE features (Muhamed et al., 2025). Affine concept editing (ACE) combines projection, a reference offset, and activation addition to control behavior (Marshall et al., 2024). SenseClamp applies a reference-based affine intervention to the paired-interpretation task: paired fit contexts determine its direction and layer-specific references, and interpretation routing selects when to intervene.

# 3 PRELIMINARIES

## 3.1 NOTATION

An *interpretation* of a surface form is a lexical sense for a common word or an entity referent for a proper name; we write *sense* as shorthand for a registered interpretation.

A sense-level unlearning request is a tuple $r = (w, s^-, s^+, \mathcal{A}_w)$. Here $w$ is the canonical surface form, and $s^-$ and $s^+$ are its distinct registered target and retained interpretations. The declared scope $\mathcal{A}_w$ is the finite set consisting of $w$ and its registered aliases.

## 3.2 PROBLEM FORMULATION

We study behavioral unlearning: success is measured by changes in answers and their likelihoods. The objective is to suppress answers about $s^-$ when a declared form is mentioned, preserve correct answers about $s^+$, and preserve behavior on queries containing no declared form.

A trusted deployer registers each request, supplying the paired interpretations, declared forms and fit contexts, and serves one request at a time; queries are single-turn inputs received through a controlled interface. In the core paired evaluation, each probe asks about one registered interpretation and contains the canonical surface form $w$; target and retained probes are posed separately and differ in their surrounding context.

The behavioral objective permits selective refusal. We report target-answer flips and retained-interpretation damage separately, and use substantive suppression as a diagnostic for losing a previously correct target answer without refusal, degeneration or expression of the queried fact anywhere in the output, together with lower reference-answer likelihood than under the unmodified model. This distinction supports comparison of response policies under the same paired registration.

# 4 SENSECLAMP

SenseClamp uses the paired registration to define both query selection and the activation edit. It controls inference in a frozen $L$-block model through hidden-state hooks. Figure 2 shows three stages: offline request registration, then online query routing and clamping of selected queries.

**Stage 1: Paired-interpretation registration.** A request supplies fit contexts for each of its two interpretations, each mentioning $w$. Let $h_{j,p}(x) \in \mathbb{R}^d$ be the unmodified model's hidden state at block $j$'s output and token position $p$ in context $x$, where $d$ is the hidden width. The target and retained sets $D^-_{\text{used}}$ and $D^+_{\text{used}}$ contain fit contexts with at least one located mention token, and

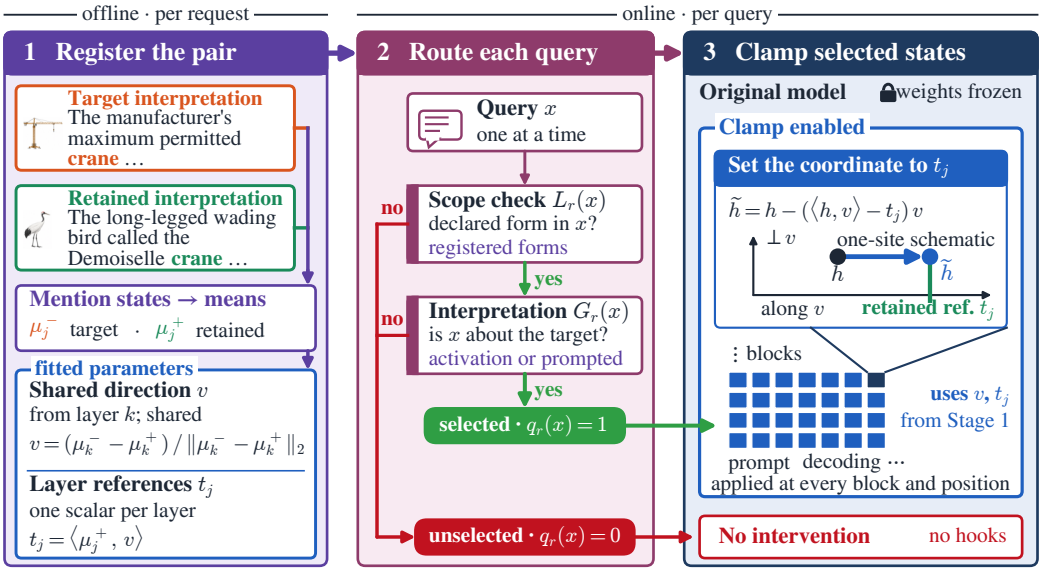



Figure 2: SenseClamp fits a shared contrast direction and layer-specific retained references (Stage 1), selects queries by lexical scope and interpretation (Stage 2), and clamps their hidden states in a frozen model (Stage 3). The cards show example fit contexts; the inset shows a single-site edit.

$P_w(x)$ contains all token positions returned by the mention locator for $x$. To give each context equal weight, we first average within each context and then across contexts:

$$\mu_j^{\pm} = \frac{1}{|D^{\pm}_{\text{used}}|} \sum_{x \in D^{\pm}_{\text{used}}} \frac{1}{|P_w(x)|} \sum_{p \in P_w(x)} h_{j,p}(x). \tag{1}$$

Their difference at one designated middle layer $k$ fixes a single unit direction shared by all layers, and the retained mean sets one scalar reference coordinate per layer:

$$v = \frac{\mu_k^- - \mu_k^+}{\|\mu_k^- - \mu_k^+\|_2}, \qquad t_j = \langle \mu_j^+, v \rangle \quad \text{for } j = 1, \ldots, L. \tag{2}$$

The contrastive direction follows representation-engineering methods (Park et al., 2024; Arditi et al., 2024; Turner et al., 2023). Under causal attention, mention states represent the prefix through the mention, so fit contexts place the distinguishing information before it. Appendix E.2 gives the locator, accepted-context counts and fit checks.

**Stage 2: Context-conditioned routing.** For a query $x$, routing combines a lexical scope check with an interpretation check, because both interpretations share the same surface form. The controller is the conjunction

$$q_r(x) = L_r(x)\, G_r(x), \qquad L_r(x),\, G_r(x),\, q_r(x) \in \{0, 1\}, \tag{3}$$

where $L_r(x)$ is a case-folded, word-boundary test for a declared surface form in $x$, and $G_r(x)$ judges whether the mention refers to $s^-$. The discriminator has two variants.

The *activation* variant uses layer $c = k$, with direction $v_{\text{cls}} = (\mu_c^- - \mu_c^+)/\|\mu_c^- - \mu_c^+\|_2$ and midpoint threshold

$$\tau = \tfrac{1}{2}\big(\langle \mu_c^-, v_{\text{cls}} \rangle + \langle \mu_c^+, v_{\text{cls}} \rangle\big), \qquad G_r(x) = \mathbf{1}\{\langle \bar{h}_c(x), v_{\text{cls}} \rangle > \tau\}. \tag{4}$$

Here $\mathbf{1}\{\cdot\}$ equals 1 when its condition is true and 0 otherwise. The query representation $\bar{h}_c(x)$ averages the unmodified query's states at its located mention positions, using the prefix available at each position; token-location and fallback rules are in Appendix E.2. The *prompted* variant presents the complete question and both interpretation glosses under labels A and B, asking for one letter. It runs two classification forwards, swapping which interpretation has which label. For order $o \in \{1, 2\}$, let $m_o(x)$ be the target-label logit minus the retained-label logit at the final prompt position.

Positive margins favour the target interpretation under either order, and $G_r(x) = \mathbf{1}\{(m_1(x) + m_2(x))/2 > 0\}$. The discriminator variant is selected once per model family on development data and held fixed during evaluation. One routing decision enables clamping for the query; generation uses the original $x$, and the classification prompt never enters it.

**Stage 3: Retained-reference clamping.** The retained fit contexts determine a scalar reference $t_j$ at each layer; these values parameterize the intervention while generation proceeds from the original query. Write $h_{j,p}$ for the state immediately before the clamp at block $j$'s output and position $p$. When $q_r(x) = 1$, we apply the update at every block, at all prompt positions during prefill and at each generated position during decoding:

$$\widetilde{h}_{j,p} \;=\; h_{j,p} - q_r(x)\left(\langle h_{j,p}, v\rangle - t_j\right) v, \qquad j = 1, \ldots, L; \tag{5}$$

when $q_r(x) = 0$, the query takes the non-intervened forward path. For a selected query, this update is a rank-one affine projection with the reference-offset form used in affine concept editing (Marshall et al., 2024):

$$\mathcal{R}_j(h) = \underset{z:\,\langle z,v\rangle = t_j}{\arg\min} \tfrac{1}{2}\|z - h\|_2^2 = h - (\langle h, v\rangle - t_j)v. \tag{6}$$

For unit $v$, the update sets its coordinate to $t_j$ and preserves the orthogonal components (Figure 2, inset). Our reference substitutions test how the choice of this parameter affects suppression and preservation.

## 5 Experimental Setup

**Datasets and models.** SensePairs-Test is the main evaluation: 20 registered target–retained interpretation pairs. SensePairs-Ext is a second, held-out evaluation of 27 new pairs, 15 lexical-sense pairs and 12 co-named-entity pairs, registered after SensePairs-Test was closed. In both sets each side of a pair has 5 single-answer factual probes, 200 and 270 probes in all, and every probe on both sides contains the same canonical surface form, so suppression and preservation are evaluated under the same required mention. This controlled benchmark tests the paired requirement of sense-level unlearning; retention on other items is measured separately. All probes underwent web fact-checking and individual human review under a fixed construction and selection protocol; the final sets were frozen before evaluation (Appendix D).

We evaluate three instruction-tuned models, Gemma-2-2B-it, Llama-3.1-8B-Instruct, and Qwen2.5-7B-Instruct (Gemma Team, Google DeepMind, 2024; Grattafiori et al., 2024; Qwen Team, 2024).

**Baselines.** Unless stated otherwise, baselines are evaluated on both main sets. **Gated prompting** places a fixed instruction with both glosses and three demonstrations per side behind the same lexical scope check as SenseClamp; SensePairs-Test only. **Activation-gated refusal** returns a fixed refusal on the queries the activation discriminator selects and the base model's output otherwise. **Partial-refusal prompting** is an instruction-only policy behind the same scope check that permits selective refusal; SensePairs-Ext only. **Same-route non-abstaining instruction** replays SenseClamp's recorded routing decisions, prepends a fixed instruction with the registered interpretations to selected questions, and reuses SenseClamp's output on unselected questions. **GUARD-IT** (Turani et al., 2026) routes queries by text cosine similarity to forget-set cluster centroids and applies a norm-preserving activation rotation to the routed ones. **LUNAR** (Shen et al., 2025) updates one selected down-projection matrix per request. **ECO** (Liu et al., 2024) corrupts the input embeddings of queries its prompt classifier flags; SensePairs-Test only. **NPO**, negative preference optimization (Zhang et al., 2024), fine-tunes per request with a negative-preference loss on the target material; Gemma-2-2B on SensePairs-Test only. **RMU**, Representation Misdirection for Unlearning (Li et al., 2024), fine-tunes per request to misdirect the intermediate representations of the target material; Gemma-2-2B on SensePairs-Test only (Appendix G.2). **DSG** (Muhamed et al., 2025) uses its own dynamic SAE classifier and selected-feature clamping within the shared lexical scope; we evaluate a development-selected adaptation on Gemma-2-2B and SensePairs-Test (Appendix B.6). LUNAR, NPO, and RMU update the language model's weights; the other systems leave those weights unchanged. All systems are evaluated on the same registered requests, all fittable methods have access to the same request-specific fit split, and fit and development probes are disjoint from evaluation probes. Table 31 (Appendix E.1) lists each method's recorded inputs, external material, and fixed or development-selected configuration.

**Configuration.** We use the midpoint-layer rule $k = \lfloor (L-1)/2 \rfloor + 1$ and unit clamp scale. The selected discriminator is activation-based for Gemma and prompted for Llama and Qwen. SensePairs-Test and SensePairs-Ext use greedy decoding with a budget of 32 generated tokens.

**Metrics.** Human metrics use *base-correct* probes, those the unmodified model answers correctly under human annotation. Within each evaluation set and model family, the base-output labels determine a common eligible probe set for all systems. Target-answer flips $F^H_{\text{tgt}}$ and retained-side damage $D^H_{\text{ret}}$ are the fractions of eligible target-side and retained-side probes subsequently answered incorrectly; flips include refusals. Substantive target suppression $S^H_{\text{fact}}$ uses the same target denominator and counts an incorrect answer with no refusal or degeneration, no expression of the queried correct fact anywhere in the output, and a strictly higher logged mean per-token reference-answer negative log-likelihood (NLL) than under the base model. The NLL values come from the original runs. Damage to other items, $C_{\text{cross}}$, is measured on base-correct probes of every other item, both sides, while the request is active; its label source is identified with each comparison. Each rate is computed within a request and averaged equally over requests with eligible probes; requests with no eligible probe are recorded as unsupported.

**Scoring and validation.** Two annotators independently labelled anonymized outputs, including base outputs, and disagreements or unsure judgments underwent blind adjudication. Appendices C.1 and B.6 report annotation coverage, agreement and adjudication for the core evaluation and additional controls.

For the main-text human evaluation, unclear labels are set to no; sensitivity analyses set them to yes, including the labels determining base-correct eligibility. We report 95% confidence intervals from a paired hierarchical bootstrap over requests, then probes. In result figures and tables, rates are percentages, differences are percentage points, and displayed intervals are 95% confidence intervals unless stated otherwise. Method contrasts subtract the comparator from SenseClamp; dashes mark unavailable configurations or measurements. The original automatic scoring protocol, its human analogue and the auxiliary LLM judge are documented in Appendix C.

# 6 Results

## 6.1 Damage to the retained interpretation

Direct evaluation of the retained interpretation reveals damage that other-item retention misses. We compare the two rates over the same requests with support in both probe pools; for the activation-gated refusal and prompting controls, both rates use human labels. On SensePairs-Test, the paired difference $D^H_{\text{ret}} - C_{\text{cross}}$ is above zero with intervals excluding zero for both controls in all three families, while $C_{\text{cross}}$ stays below one percentage point. For example, Qwen2.5-7B gated prompting damages 0.2% of other-item answers and 20.2% of answers about the retained interpretation. The same-probe check using the original automatic scoring supports these contrasts when each retained question is held fixed across requests (Appendix A.4.1). On SensePairs-Ext, partial-refusal prompting likewise causes greater damage to the retained interpretation than to other items in all three families, with paired intervals excluding zero with unclear labels set to either no or yes.

## 6.2 Human-evaluated suppression and preservation

Across all three families, SenseClamp achieves higher human-verified substantive suppression than LUNAR on both sets, gated prompting on Test, and partial-refusal prompting on Ext, with every paired interval excluding zero (Figure 3; Appendix A). It also improves suppression over GUARD-IT on Llama and Qwen in both sets.

On SensePairs-Test, SenseClamp achieves higher suppression and lower retained-side damage than LUNAR and gated prompting on Llama and Qwen. Damage reductions are 22.0–37.8 points against LUNAR and 9.7–18.9 against prompting, with all paired intervals excluding zero (Appendix A.2). On SensePairs-Ext, SenseClamp has higher suppression and lower damage than these comparators and GUARD-IT on Llama and Qwen under the primary human-label resolution. Resolving unclear labels as yes preserves the suppression conclusions, although the Llama damage comparison with LUNAR then includes zero (95% CI -17.6 to 0.2 points). For Gemma on the extension set, damage

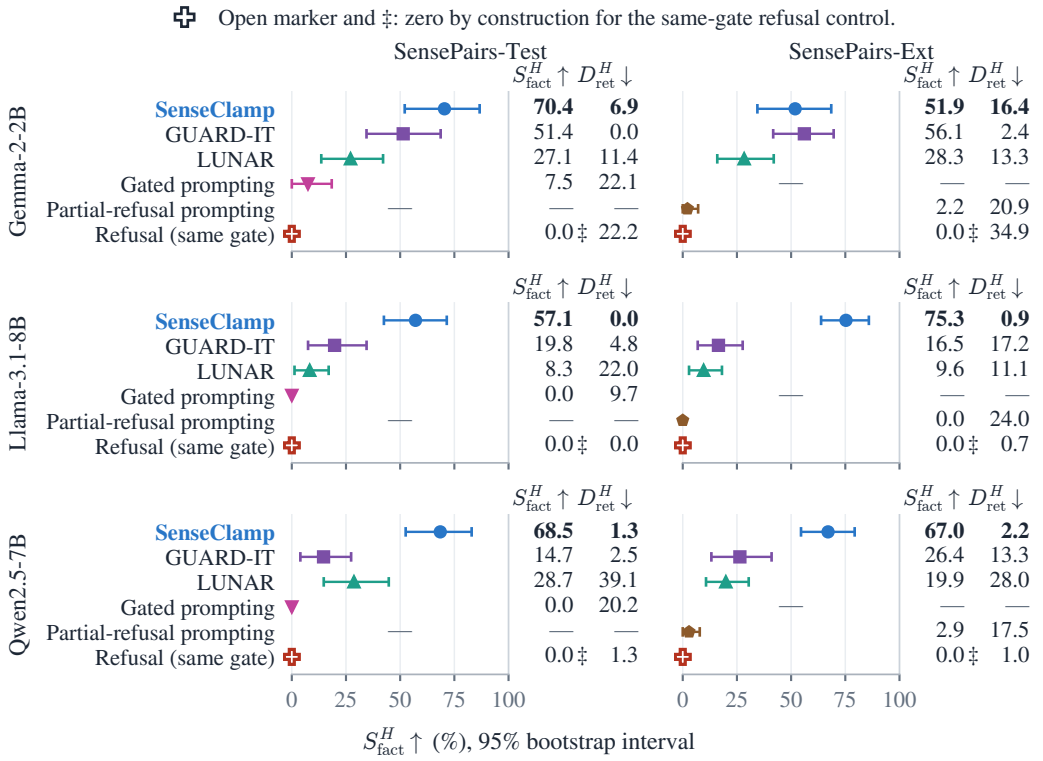



Figure 3: Human evaluation compares target suppression and retained-interpretation damage across three model families. SenseClamp achieves higher substantive suppression than LUNAR on both evaluation sets, gated prompting on SensePairs-Test, and partial-refusal prompting on SensePairs-Ext.

Table 1: This table compares SenseClamp with a same-route instruction and a DSG adaptation. SenseClamp achieves stronger substantive suppression than both controls.

| Set | Model | Method | $F^H_{\text{tgt}} \uparrow$ | $S^H_{\text{fact}} \uparrow$ | $D^H_{\text{ret}} \downarrow$ |
|---|---|---|---|---|---|
| Test | Gemma-2-2B | SenseClamp | 86.8 | 70.4 | 6.9 |
| | | Same-route instruction | 3.1 | 3.1 | 1.4 |
| | | DSG adaptation | 90.4 | 7.0 | 38.3 |
| Test | Llama-3.1-8B | SenseClamp | 71.3 | 57.1 | 0.0 |
| | | Same-route instruction | 8.7 | 5.4 | 0.0 |
| Test | Qwen2.5-7B | SenseClamp | 80.2 | 68.5 | 1.3 |
| | | Same-route instruction | 11.6 | 9.2 | 0.0 |
| Ext | Gemma-2-2B | SenseClamp | 71.7 | 51.9 | 16.4 |
| | | Same-route instruction | 7.2 | 1.9 | 0.0 |
| Ext | Llama-3.1-8B | SenseClamp | 91.4 | 75.3 | 0.9 |
| | | Same-route instruction | 4.1 | 4.1 | 0.9 |
| Ext | Qwen2.5-7B | SenseClamp | 87.1 | 67.0 | 2.2 |
| | | Same-route instruction | 11.4 | 6.8 | 1.3 |

exceeds GUARD-IT by +14.0 points (2.3 to 28.3). The extension uses the original layer rule, clamp scale and family-specific discriminator variant, fitting the direction and references from each new pair's own contexts.

**Dynamic SAE clamping.** On Gemma-2-2B, the DSG adaptation flips 90.4% of eligible target answers, with 87.3% degeneration, 7.0% substantive suppression and 38.3% retained-side damage

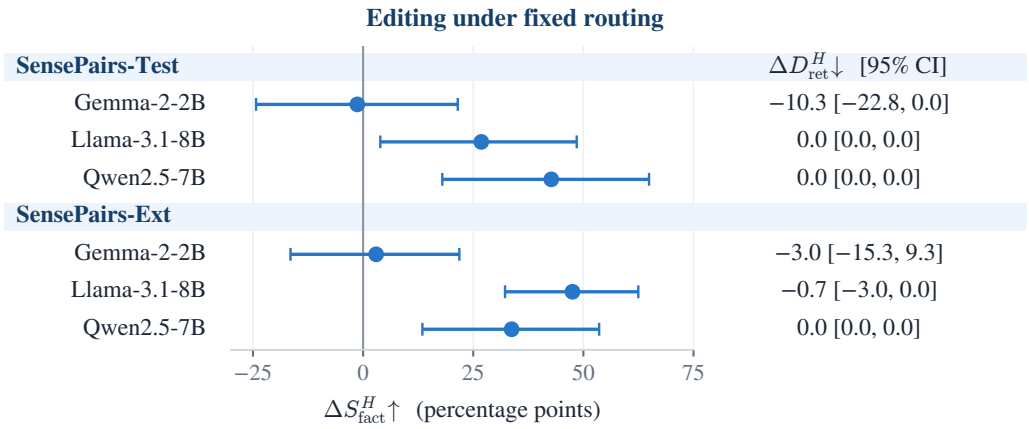



Figure 4: Human evaluation compares SenseClamp's edit with GUARD-IT's rotation under identical routing. SenseClamp improves substantive suppression on Llama and Qwen across both sets.

(Table 1). The development fallback selects 20 features after all four candidates exceed the retained-damage threshold. Appendix B.6 specifies the shared settings and feature-selection differences from the published configuration (Muhamed et al., 2025).

**Response policy.** A same-gate control replaces every selected answer with a fixed refusal, enabling comparison through target-answer flips $F^H_{\text{tgt}}$ and damage $D^H_{\text{ret}}$ (Table 7). Refusal flips more target answers on Llama and Qwen in both sets, while clamping causes less retained-side damage on Gemma. Among 31 human-base-correct retained probes selected on the Gemma extension set, clamping damages 15, compared with all 31 under refusal. On Gemma, preservation therefore depends on the response strategy and routing.

**Output content.** Among human-verified successes, the auxiliary judge labels most SenseClamp responses as unrelated to either interpretation, with a smaller share consistent with the retained interpretation (Appendix C.5).

**Scoring robustness.** Partial-refusal prompting on SensePairs-Ext with Llama scores 96.3% under the original automatic suppression metric but 0.0% under human substantive suppression. The human diagnostic changes labels, adds fact absence and recomputes eligibility, using the same outputs and NLLs. Human suppression conclusions persist without fact absence and under either unclear-label resolution (Appendix C.2; Figure 5).

### 6.3 Contributions of Routing and Reference Clamping

**Instruction under identical routing.** Under SenseClamp's recorded routing, a fixed instruction to suppress target facts without refusing flips only 3.1–11.6% of eligible target answers across the six model–set combinations, versus 71.3–91.4% for SenseClamp (Table 1). SenseClamp's substantive suppression is 50.0–71.2 points higher, with all paired intervals above zero. On Gemma Ext, the instruction has 0.0% retained-side damage versus 16.4% for SenseClamp (difference 16.4, 95% CI 5.3 to 29.9). Suppression gains persist under the alternative label resolution and after excluding 45 cells with answer-string overlap in their instructions (Appendix B.6).

**Editing under fixed routing.** With identical recorded routing, SenseClamp's editing configuration improves human substantive suppression over GUARD-IT's rotation on Llama and Qwen in both sets (Figure 4). The comparison concerns the complete editing configuration, including sites and strength; complementary routing controls appear in Appendices B.1 and B.5.

**Retained-reference substitutions.** On Gemma, the retained reference improves human suppression over zero and away-from-retained destinations at their original scales (Table 2). The suppres-

Table 2: Reference substitutions test how the retained reference contributes to SenseClamp. The retained reference improves suppression over zero references on Gemma and Qwen and over donor references on Qwen.

| **Configuration** | $\alpha$ | $\Delta S^H_{\text{fact}}$ **[95% CI]** | $\Delta D^H_{\text{ret}}$ **[95% CI]** |
|---|---|---|---|
| *A. Destination, Gemma-2-2B* | | | |
| Zero reference *(original scale)* | 1 | +40.9 [19.7, 61.8] | +4.2 [-5.6, 15.3] |
| Zero reference *(calibrated)* | 2 | +29.7 [4.0, 55.3] | +4.2 [-5.6, 15.3] |
| Away from retained *(original scale)* | 1 | +70.4 [52.1, 86.7] | -15.3 [-31.9, -2.8] |
| Away from retained *(calibrated)* | 0.1 | +7.9 [-15.6, 30.6] | -15.3 [-31.9, -2.8] |
| *B. Reference value, Qwen2.5-7B* | | | |
| Zero reference *(original scale)* | 1 | +48.2 [29.3, 66.1] | +1.3 [0.0, 5.0] |
| Zero reference *(calibrated)* | 2 | +44.2 [24.0, 63.3] | +1.3 [0.0, 5.0] |
| Donor reference 0 *(approximated)* | 2 | +49.8 [29.2, 69.0] | +1.3 [0.0, 5.0] |
| Donor reference 1 *(calibrated)* | 2 | +50.7 [35.3, 66.3] | +1.3 [0.0, 5.0] |
| Donor reference 2 *(approximated)* | 2 | +49.1 [31.5, 66.2] | +1.3 [0.0, 5.0] |

sion gain over the zero reference persists after development-set displacement calibration. On Qwen, it improves suppression by +44.2–+50.7 points over zero and references from other requests, holding routing, direction and sites fixed. All five suppression intervals exclude zero under both label scenarios. Selected scales match decoding displacement within a preset tolerance where feasible (*calibrated*), otherwise using the closest grid value (*approximated*); Appendix B.2.2 gives calibration and donor controls. In Table 2, $\alpha$ gives each control's scale.

**Direction controls.** On Gemma, the fitted direction improves human suppression over three-seed random-direction controls at both original and development-selected scales; residual displacement differences are reported in Appendix B.3.

### 6.4 Context sensitivity and scope

**Effect of disambiguating context.** We evaluate 188 cue-before and cue-after rewrites of 94 facts covering all 47 registered pairs. Routing is evaluated on all requests, while answer outcomes are compared on probes that the base model answers correctly in both versions. Paired rewrites show that placing the cue after the mention reduces target selection and increases retained-side selection for Gemma's activation discriminator. On the common base-correct facts, human suppression falls by 31.4 points (95% CI 14.3 to 48.6), while retained-side damage rises by 22.6 points (9.7 to 38.7). In the router-substitution diagnostic on Gemma, full-question routing reduces extension-set damage from 16.4% to 0.0%. Qwen's suppression varies across rewrites with identical routing, indicating response sensitivity after query selection (Appendix B.4).

**Scope beyond separately posed queries.** SensePairs-Stress is a supplementary evaluation of mixed queries, queries about unregistered interpretations, and alias-free queries. Its alias-free queries bypass the lexical gate and retain the target answer on all automatically base-correct probes: 38/38 for Gemma, 48/48 for Llama and 40/40 for Qwen. SensePairs-Joint requires target-fact suppression and a correct retained answer in the same response. Its calibrated LLM judge gives Llama joint success of 23.7% for SenseClamp and 78.1% for partial-refusal prompting, a paired difference favoring prompting of +54.4 points (95% CI 39.5 to 69.3). These boundary evaluations identify alias coverage and mixed-response control as distinct requirements beyond the registered, separately posed task (Appendices F.1 and F.2).

## 7 Discussion and Conclusion

Paired human evaluation reveals retained-interpretation damage missed by other-item retention and distinguishes refusal, degeneration and well-formed, non-abstaining responses omitting the queried fact. SenseClamp's controls show how its editing configuration and retained reference contribute to target suppression. Future work addresses fuller context, aliases and mixed requests.

## Reproducibility Statement

The supplementary archive provides the SenseClamp implementation, the SensePairs benchmark, and the configurations and instructions needed to run the method. It contains method code and benchmark data; experiment outputs, human annotation records, aggregate results, and baseline implementations are outside the release. Appendix B describes the component and post hoc control protocols, Appendix E documents implementation settings, and Appendix C.3 records the scoring revision.

## Ethics Statement

SensePairs contains questions about publicly documented entities and word senses, drawn from public sources and individually reviewed; no private personal data is collected for benchmark construction. Two annotators independently evaluate model outputs under a blinded protocol, and disagreements underwent blind adjudication (Appendix C.1). We assess the need for ethics approval for this annotation procedure and considered external approval unnecessary for this limited evaluation of model outputs concerning publicly documented facts.

## Use of Large Language Models

Large language models assisted in polishing the manuscript; for example, by checking for grammar, spelling, and display errors. LLMs generated candidate benchmark questions following the SensePairs construction protocol. These candidates underwent automated validation, web fact-checking, and individual human review before selection for the benchmark (Appendix D). The auxiliary output judge uses a fixed commercial LLM snapshot at temperature 0 with a rubric fixed before any call (Appendix C.5). The main human evaluation used two independent annotators and blind adjudication. The authors take responsibility for the final text, code, and reported findings.

## References

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 (NeurIPS)*, 2024.

Noam Diamant, Neta Glazer, and Ethan Fetaya. Stress testing unlearning algorithms. *arXiv preprint arXiv:2608.22527*, 2026. URL https://arxiv.org/abs/2608.22527.

Gemma Team, Google DeepMind. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Shengyuan Hu, Neil Kale, Pratiksha Thaker, Yiwei Fu, Steven Wu, and Virginia Smith. BLUR: A benchmark for LLM unlearning robust to forget-retain overlap, 2025. URL https://arxiv.org/abs/2506.15699.

Wonje Jeung, Sangyeon Yoon, Hyesoo Hong, Soeun Kim, Seungju Han, Youngjae Yu, and Albert No. DUSK: Do not unlearn shared knowledge. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. URL https://aclanthology.org/2026.findings-acl.2085/.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. RWKU: Benchmarking real-world knowledge unlearning for large language models. *arXiv preprint arXiv:2406.10890*, 2024.

Sahil Kale and Ian Harris. ConceptGuard: Benchmarking context-sensitive unlearning in large language models, 2026. URL https://arxiv.org/abs/2608.20338.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2409.05907.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning (ICML)*, pp. 28525–28550, 2024.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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 (COLM)*, 2024.

Thomas Marshall, Adam Scherlis, and Nora Belrose. Refusal in LLMs is an affine function. *arXiv preprint arXiv:2411.09003*, 2024. URL https://arxiv.org/abs/2411.09003.

Aashiq Muhamed, Jacopo Bonato, Mona Diab, and Virginia Smith. SAEs can improve unlearning: Dynamic sparse autoencoder guardrails for precision unlearning in LLMs. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=kaPAalWAp3.

Office of the Privacy Commissioner of Canada, Commission d'accès à l'information du Québec, Office of the Information and Privacy Commissioner for British Columbia, and Office of the Information and Privacy Commissioner of Alberta. PIPEDA findings #2026-002: Joint investigation of OpenAI OpCo, LLC. Joint report of findings, released May 6, 2026, 2026. URL https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2026/pipeda-2026-002/.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *International Conference on Machine Learning (ICML)*, 2024.

Amit Peleg, Naman Deep Singh, Naama Pearl, Bibhabasu Mohapatra, and Matthias Hein. Forget narrowly, retain broadly: Unlearning as an asymmetric generalization problem. *arXiv preprint arXiv:2607.09236*, 2026.

Yuefeng Peng, Parnian Afshar, Megan Ganji, Thomas Butler, Amir Houmansadr, Mingxian Wang, and Dezhi Hong. Forget to know, remember to use: Context-aware unlearning for large language models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2510.17620.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

William A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325, 1955. doi: 10.1086/266577.

William Shen, Xinchi Qiu, Meghdad Kurmanji, Alexandru-Andrei Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D. Lane. LLM unlearning via neural activation redirection. In *Advances in Neural Information Processing Systems*, volume 38, pp. 44253–44290, 2025. doi: 10.52202/085713-1475. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3ed6a903be20e1478aa6793f14938091-Paper-Conference.pdf.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Minkyoo Song, Hanna Kim, Jaehan Kim, Seungwon Shin, and Sooel Son. Refusal is not an option: Unlearning safety alignment of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 319–338, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/song-minkyoo.

Shota Takashiro, Takeshi Kojima, Andrew Gambardella, Qi Cao, Yusuke Iwasawa, and Yutaka Matsuo. Answer when needed, forget when not: Language models pretend to forget via in-context knowledge unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24872–24885, 2025. doi: 10.18653/v1/2025.findings-acl.1276. URL https://aclanthology.org/2025.findings-acl.1276/.

Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. To forget or not? towards practical knowledge unlearning for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1524–1537. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.82. URL https://aclanthology.org/2024.findings-emnlp.82/.

Vinícius Conte Turani, Otávio Parraga, João Vitor Boer Abitante, Kristen K Arguello, Joana Pasquali, Ramiro N Barros, Flavio du Pin Calmon, Christian Mattjie, Rodrigo C Barros, and Lucas S Kupssinskü. Inference-time machine unlearning via gated activation redirection. *arXiv preprint arXiv:2605.12765*, 2026.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Yaxuan Wang, Chris Yuhao Liu, Quan Liu, Jinlong Pang, Wei Wei, Yujia Bao, and Yang Liu. DRAGON: Guard LLM unlearning in context via negative detection and reasoning. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=vQLUAkl5SG.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling (COLM)*, 2024.

# A FULL HUMAN-EVALUATED RESULTS

## A.1 ABSOLUTE RESULTS AND SUPPORT

Tables 3 and 4 collect the complete primary human estimates, including the same-route instruction and Gemma Test DSG controls. Each family uses one human base-correct probe set for all methods; requests without eligible probes do not enter the corresponding rate. The prompting comparator is gated prompting on Test and partial-refusal prompting on Ext. Method definitions are in Appendix B.6; the shared annotation and estimation protocol is in Appendix C.1. The primary tables set unclear labels to no and report percentages with 95% percentile bootstrap intervals; $N/P$ denotes supported requests and eligible probes separately for each side. Same-gate refusal has zero $S^H_{\mathrm{fact}}$ by construction.

Table 3: This table reports human suppression and retained-side damage on SensePairs-Test. SenseClamp achieves high non-abstaining suppression with low retained-side damage across the three families.

| System | $F^H_{\mathrm{tgt}}$ [95% CI] | $S^H_{\mathrm{fact}}$ [95% CI] | $D^H_{\mathrm{ret}}$ [95% CI] |
|---|---|---|---|
| *Gemma-2-2B;* $N_t/P_t = 19/66$, $N_r/P_r = 18/56$ | | | |
| **SenseClamp** | 86.8 [71.1, 100.0] | 70.4 [52.1, 86.7] | 6.9 [0.0, 16.7] |
| GUARD-IT | 71.8 [57.4, 84.8] | 51.4 [34.5, 68.7] | 0.0 [0.0, 0.0] |
| LUNAR | 67.6 [50.9, 82.5] | 27.1 [13.6, 42.1] | 11.4 [1.4, 23.9] |
| Gated prompting | 20.2 [7.0, 35.1] | 7.5 [0.0, 18.4] | 22.1 [6.9, 39.8] |
| Same-gate refusal | 100.0 [100.0, 100.0] | 0.0 [0.0, 0.0]‡ | 22.2 [5.6, 44.4] |
| Same-route instruction | 3.1 [0.0, 10.1] | 3.1 [0.0, 10.1] | 1.4 [0.0, 5.6] |
| DSG adaptation | 90.4 [80.3, 98.3] | 7.0 [0.0, 17.5] | 38.3 [20.8, 56.7] |
| *Llama-3.1-8B;* $N_t/P_t = 20/90$, $N_r/P_r = 20/82$ | | | |
| **SenseClamp** | 71.3 [57.0, 84.8] | 57.1 [42.5, 71.5] | 0.0 [0.0, 0.0] |
| GUARD-IT | 73.3 [60.8, 85.3] | 19.8 [7.5, 34.5] | 4.8 [0.0, 12.5] |
| LUNAR | 77.8 [64.7, 89.6] | 8.3 [1.3, 17.0] | 22.0 [10.6, 34.8] |
| Gated prompting | 94.5 [87.0, 100.0] | 0.0 [0.0, 0.0] | 9.7 [2.3, 18.9] |
| Same-gate refusal | 94.5 [87.0, 100.0] | 0.0 [0.0, 0.0]‡ | 0.0 [0.0, 0.0] |
| Same-route instruction | 8.7 [2.0, 17.1] | 5.4 [0.0, 12.9] | 0.0 [0.0, 0.0] |
| *Qwen2.5-7B;* $N_t/P_t = 19/69$, $N_r/P_r = 20/71$ | | | |
| **SenseClamp** | 80.2 [64.3, 93.3] | 68.5 [52.5, 83.0] | 1.3 [0.0, 5.0] |
| GUARD-IT | 66.8 [50.6, 81.8] | 14.7 [4.0, 27.4] | 2.5 [0.0, 8.8] |
| LUNAR | 76.0 [60.4, 89.4] | 28.7 [14.7, 44.7] | 39.1 [22.8, 55.8] |
| Gated prompting | 100.0 [100.0, 100.0] | 0.0 [0.0, 0.0] | 20.2 [6.7, 35.8] |
| Same-gate refusal | 98.3 [93.0, 100.0] | 0.0 [0.0, 0.0]‡ | 1.3 [0.0, 5.0] |
| Same-route instruction | 11.6 [1.3, 25.0] | 9.2 [0.0, 22.4] | 0.0 [0.0, 0.0] |

## A.2 PAIRED COMPARISONS AND LABEL SENSITIVITY

Table 5 reports SenseClamp minus each comparator using the stored paired estimates and intervals. Positive target-flip and suppression differences favour SenseClamp; negative damage differences favour SenseClamp. The primary scenario sets unclear labels to no, and the alternative sets them to yes throughout, including base eligibility; each scenario reports its own support. Intervals resample requests and then eligible probes, using 5,000 paired draws with seed 0 and no multiplicity adjustment. The alternative scenario preserves all six suppression gains over the instruction control, both the suppression and damage advantages over DSG, and the Gemma Ext damage cost relative to the instruction control. Appendix C.2 explains the core comparisons whose interval status changes.

Table 4: This table reports human results on the new interpretation pairs in SensePairs-Ext. SenseClamp combines stronger substantive suppression with lower retained-side damage than GUARD-IT, LUNAR and partial-refusal prompting on Llama and Qwen.

| System | $F^H_{\text{tgt}}$ [95% CI] | $S^H_{\text{fact}}$ [95% CI] | $D^H_{\text{ret}}$ [95% CI] |
|---|---|---|---|
| *Gemma-2-2B;* $N_t/P_t = 27/88$, $N_r/P_r = 25/91$ | | | |
| **SenseClamp** | 71.7 [55.0, 87.4] | 51.9 [34.4, 68.5] | 16.4 [5.3, 29.9] |
| GUARD-IT | 81.8 [71.4, 90.7] | 56.1 [41.7, 69.6] | 2.4 [0.0, 6.4] |
| LUNAR | 61.4 [47.6, 74.8] | 28.3 [15.9, 42.0] | 13.3 [4.0, 24.7] |
| Partial-refusal prompting | 35.6 [23.0, 49.1] | 2.2 [0.0, 7.1] | 20.9 [10.0, 33.1] |
| Same-gate refusal | 73.2 [55.6, 88.9] | 0.0 [0.0, 0.0]‡ | 34.9 [16.8, 54.1] |
| Same-route instruction | 7.2 [0.9, 15.6] | 1.9 [0.0, 7.4] | 0.0 [0.0, 0.0] |
| *Llama-3.1-8B;* $N_t/P_t = 27/112$, $N_r/P_r = 27/102$ | | | |
| **SenseClamp** | 91.4 [83.2, 97.8] | 75.3 [63.8, 85.9] | 0.9 [0.0, 3.7] |
| GUARD-IT | 82.3 [70.9, 92.2] | 16.5 [6.9, 27.7] | 17.2 [7.8, 28.2] |
| LUNAR | 77.0 [66.9, 86.5] | 9.6 [2.9, 18.1] | 11.1 [3.0, 22.0] |
| Partial-refusal prompting | 97.8 [94.1, 100.0] | 0.0 [0.0, 0.0] | 24.0 [11.7, 38.0] |
| Same-gate refusal | 97.0 [91.1, 100.0] | 0.0 [0.0, 0.0]‡ | 0.7 [0.0, 3.0] |
| Same-route instruction | 4.1 [0.0, 10.0] | 4.1 [0.0, 10.0] | 0.9 [0.0, 3.7] |
| *Qwen2.5-7B;* $N_t/P_t = 27/99$, $N_r/P_r = 26/91$ | | | |
| **SenseClamp** | 87.1 [77.4, 95.4] | 67.0 [54.6, 79.3] | 2.2 [0.0, 7.1] |
| GUARD-IT | 77.9 [65.7, 88.6] | 26.4 [13.2, 41.0] | 13.3 [5.4, 22.6] |
| LUNAR | 88.6 [80.1, 95.7] | 19.9 [10.7, 30.4] | 28.0 [16.0, 41.7] |
| Partial-refusal prompting | 86.7 [77.4, 94.5] | 2.9 [0.0, 7.8] | 17.5 [7.3, 29.2] |
| Same-gate refusal | 98.8 [95.1, 100.0] | 0.0 [0.0, 0.0]‡ | 1.0 [0.0, 3.9] |
| Same-route instruction | 11.4 [3.7, 21.0] | 6.8 [0.9, 14.2] | 1.3 [0.0, 5.1] |

Table 5: This table reports paired human differences under both unclear-label scenarios. SenseClamp's substantive-suppression gains retain the same interval status across the scenarios.

| **Comparator** | $\Delta F^H_{\text{tgt}}$ **[95% CI]** | $\Delta S^H_{\text{fact}}$ **[95% CI]** | $\Delta D^H_{\text{ret}}$ **[95% CI]** |
|---|---|---|---|
| **Unclear labels set to no** | | | |
| **SensePairs-Test** | | | |
| *Gemma-2-2B;* $N_t/P_t = 19/66$, $N_r/P_r = 18/56$ | | | |
| GUARD-IT | +15.1 [-4.8, 33.3] | +19.0 [-5.1, 41.7] | +6.9 [0.0, 16.7] |
| LUNAR | +19.2 [1.8, 37.6] | +43.3 [18.6, 65.4] | -4.5 [-19.7, 11.1] |
| Gated prompting | +66.7 [46.5, 85.1] | +63.0 [40.4, 82.0] | -15.2 [-36.1, 4.7] |
| Same-route instruction | +83.8 [66.7, 98.3] | +67.4 [48.2, 85.1] | +5.6 [0.0, 15.3] |
| DSG adaptation | -3.5 [-21.5, 12.7] | +63.4 [40.9, 83.2] | -31.4 [-50.6, -14.4] |
| *Llama-3.1-8B;* $N_t/P_t = 20/90$, $N_r/P_r = 20/82$ | | | |
| GUARD-IT | -2.0 [-15.5, 11.5] | +37.3 [16.8, 56.8] | -4.8 [-12.5, 0.0] |
| LUNAR | -6.5 [-21.8, 8.3] | +48.8 [33.5, 63.5] | -22.0 [-34.8, -10.6] |
| Gated prompting | -23.2 [-39.0, -7.4] | +57.1 [42.5, 71.5] | -9.7 [-18.9, -2.3] |
| Same-route instruction | +62.7 [46.5, 78.3] | +51.7 [37.0, 66.8] | +0.0 [0.0, 0.0] |
| *Qwen2.5-7B;* $N_t/P_t = 19/69$, $N_r/P_r = 20/71$ | | | |
| GUARD-IT | +13.4 [-8.8, 34.0] | +53.9 [32.3, 74.0] | -1.3 [-8.8, 5.0] |
| LUNAR | +4.2 [-16.4, 25.4] | +39.8 [23.1, 56.6] | -37.8 [-54.3, -22.0] |
| Gated prompting | -19.8 [-35.7, -6.7] | +68.5 [52.5, 83.0] | -18.9 [-34.5, -6.0] |
| Same-route instruction | +68.6 [49.8, 85.4] | +59.3 [41.3, 76.2] | +1.3 [0.0, 5.0] |
| **SensePairs-Ext** | | | |
| *Gemma-2-2B;* $N_t/P_t = 27/88$, $N_r/P_r = 25/91$ | | | |
| GUARD-IT | -10.1 [-29.3, 8.6] | -4.3 [-28.0, 18.5] | +14.0 [2.3, 28.3] |
| LUNAR | +10.3 [-10.8, 30.6] | +23.5 [1.5, 45.6] | +3.1 [-12.7, 20.3] |
| Partial-refusal prompting | +36.1 [13.5, 57.4] | +49.7 [30.9, 67.8] | -4.5 [-22.3, 14.3] |
| Same-route instruction | +64.5 [48.2, 80.6] | +50.0 [33.0, 66.7] | +16.4 [5.3, 29.9] |
| *Llama-3.1-8B;* $N_t/P_t = 27/112$, $N_r/P_r = 27/102$ | | | |
| GUARD-IT | +9.1 [-4.7, 23.8] | +58.8 [44.9, 72.2] | -16.3 [-27.6, -6.1] |
| LUNAR | +14.3 [4.1, 25.1] | +65.7 [52.6, 78.0] | -10.2 [-21.3, -1.5] |

*continued on next page*

Table 5: This table reports paired human differences under both unclear-label scenarios. SenseClamp's substantive-suppression gains retain the same interval status across the scenarios. (continued)

| **Comparator** | $\Delta F^H_{\text{tgt}}$ **[95% CI]** | $\Delta S^H_{\text{fact}}$ **[95% CI]** | $\Delta D^H_{\text{ret}}$ **[95% CI]** |
|---|---|---|---|
| Partial-refusal prompting | -6.4 [-15.1, 0.7] | +75.3 [63.8, 85.9] | -23.1 [-37.6, -10.2] |
| Same-route instruction | +87.2 [78.1, 94.9] | +71.2 [58.2, 82.8] | +0.0 [0.0, 0.0] |
| *Qwen2.5-7B;* $N_t/P_t = 27/99$*,* $N_r/P_r = 26/91$ | | | |
| GUARD-IT | +9.2 [-4.9, 24.2] | +40.7 [21.5, 59.4] | -11.1 [-21.2, -1.3] |
| LUNAR | -1.5 [-12.2, 9.1] | +47.2 [32.4, 62.1] | -25.8 [-39.6, -13.3] |
| Partial-refusal prompting | +0.4 [-12.5, 12.9] | +64.1 [50.6, 77.0] | -15.3 [-26.2, -5.9] |
| Same-route instruction | +75.7 [62.4, 87.3] | +60.3 [45.1, 74.5] | +1.0 [0.0, 3.9] |
| **Unclear labels set to yes** | | | |
| **SensePairs-Test** | | | |
| *Gemma-2-2B;* $N_t/P_t = 19/66$*,* $N_r/P_r = 18/56$ | | | |
| GUARD-IT | +14.0 [-6.8, 33.5] | +18.0 [-6.8, 41.2] | +6.9 [0.0, 16.7] |
| LUNAR | +23.3 [3.7, 43.8] | +46.1 [20.9, 68.9] | -4.5 [-19.7, 11.1] |
| Gated prompting | +64.6 [45.2, 82.9] | +60.9 [38.3, 80.3] | -15.2 [-36.1, 4.7] |
| Same-route instruction | +81.7 [64.3, 96.1] | +65.3 [45.6, 83.3] | +5.6 [0.0, 15.3] |
| DSG adaptation | -5.6 [-24.2, 11.1] | +61.3 [38.2, 81.2] | -31.4 [-50.6, -14.4] |
| *Llama-3.1-8B;* $N_t/P_t = 20/90$*,* $N_r/P_r = 20/82$ | | | |
| GUARD-IT | -2.0 [-15.5, 11.5] | +37.3 [16.8, 56.8] | -4.8 [-12.5, 0.0] |
| LUNAR | -4.0 [-19.5, 11.3] | +51.3 [36.0, 66.3] | -20.8 [-33.5, -9.3] |
| Gated prompting | -23.2 [-39.0, -7.4] | +57.1 [42.5, 71.5] | -9.7 [-18.9, -2.3] |
| Same-route instruction | +62.7 [46.5, 78.3] | +51.7 [37.0, 66.8] | +0.0 [0.0, 0.0] |
| *Qwen2.5-7B;* $N_t/P_t = 19/70$*,* $N_r/P_r = 20/72$ | | | |
| GUARD-IT | +12.1 [-10.3, 34.0] | +53.4 [32.4, 73.3] | -1.3 [-8.8, 3.8] |
| LUNAR | +3.3 [-17.4, 24.0] | +40.3 [24.1, 55.7] | -38.7 [-55.1, -23.1] |
| Gated prompting | -21.6 [-37.5, -7.8] | +67.6 [51.8, 82.6] | -18.9 [-34.5, -6.0] |
| Same-route instruction | +66.8 [47.9, 84.5] | +58.4 [40.0, 76.1] | +1.3 [0.0, 5.0] |
| **SensePairs-Ext** | | | |
| *Gemma-2-2B;* $N_t/P_t = 27/90$*,* $N_r/P_r = 25/93$ | | | |
| GUARD-IT | -7.5 [-26.2, 12.1] | -3.2 [-26.0, 19.7] | +14.4 [3.2, 27.7] |
| LUNAR | +17.6 [-3.3, 38.5] | +30.0 [8.5, 51.2] | +2.7 [-13.1, 19.0] |
| Partial-refusal prompting | +35.2 [12.3, 57.0] | +49.1 [30.4, 67.8] | -3.1 [-21.9, 16.1] |
| Same-route instruction | +64.5 [48.3, 80.6] | +50.0 [33.5, 67.0] | +16.0 [5.1, 28.8] |
| *Llama-3.1-8B;* $N_t/P_t = 27/113$*,* $N_r/P_r = 27/103$ | | | |
| GUARD-IT | +9.5 [-4.9, 23.7] | +59.8 [47.1, 72.0] | -15.7 [-26.7, -5.9] |
| LUNAR | +14.1 [2.8, 25.9] | +66.1 [52.9, 78.5] | -7.6 [-17.6, 0.2] |
| Partial-refusal prompting | -7.9 [-16.3, -0.7] | +73.8 [62.2, 84.5] | -12.6 [-25.4, -1.9] |
| Same-route instruction | +85.7 [76.7, 93.6] | +69.7 [56.7, 81.7] | +0.0 [0.0, 0.0] |
| *Qwen2.5-7B;* $N_t/P_t = 27/102$*,* $N_r/P_r = 26/94$ | | | |
| GUARD-IT | +9.8 [-5.3, 24.8] | +40.7 [21.9, 59.3] | -13.5 [-25.1, -3.1] |
| LUNAR | -0.1 [-10.6, 10.9] | +48.5 [33.0, 63.6] | -27.7 [-41.1, -15.1] |
| Partial-refusal prompting | +1.2 [-12.7, 14.8] | +65.3 [52.2, 78.0] | -13.2 [-23.0, -5.1] |
| Same-route instruction | +74.8 [61.2, 86.7] | +60.3 [45.2, 74.3] | +1.0 [0.0, 3.9] |

### A.3 Response types and selective refusal

Table 6 separates target-side refusal and degeneration from suppression for the additional controls and their matched SenseClamp runs. The same-gate refusal control replays SenseClamp's recorded routing, returning the fixed refusal on selected queries and the stored base output otherwise. On a human-eligible target probe, the selected branch refuses and the other branch preserves the correct base answer, so neither earns substantive-suppression credit. This derived policy requires no new generation or likelihood measurement and avoids the batching drift of SenseClamp's unmodified path (Appendix E.3). It differs from the original activation-gated refusal arm, which uses the activation discriminator in every family. Table 7 compares target flips and retained damage under both label scenarios; † marks an interval excluding zero.

Table 6: This table separates refusal and degeneration in the additional human comparisons. SenseClamp's substantive-suppression score excludes both response types.

| System | Unclear | Refusal [95% CI] | Degeneration [95% CI] |
|---|---|---|---|
| *SensePairs-Test / Gemma-2-2B* | | | |
| SenseClamp | no | 1.8 [0.0, 7.0] | 11.6 [1.1, 25.0] |
| SenseClamp | yes | 3.9 [0.0, 11.2] | 13.7 [2.1, 28.2] |
| Same-route instruction | no | 0.0 [0.0, 0.0] | 0.0 [0.0, 0.0] |
| Same-route instruction | yes | 0.0 [0.0, 0.0] | 0.0 [0.0, 0.0] |
| DSG adaptation | no | 6.1 [0.0, 15.5] | 87.3 [73.7, 97.4] |
| DSG adaptation | yes | 6.1 [0.0, 15.5] | 87.3 [73.7, 97.4] |
| *SensePairs-Test / Llama-3.1-8B* | | | |
| SenseClamp | no | 5.3 [0.0, 13.0] | 5.5 [0.0, 14.0] |
| SenseClamp | yes | 5.3 [0.0, 13.0] | 5.5 [0.0, 14.0] |
| Same-route instruction | no | 1.0 [0.0, 4.0] | 0.0 [0.0, 0.0] |
| Same-route instruction | yes | 1.0 [0.0, 4.0] | 0.0 [0.0, 0.0] |
| *SensePairs-Test / Qwen2.5-7B* | | | |
| SenseClamp | no | 5.8 [0.0, 14.0] | 4.6 [0.0, 13.7] |
| SenseClamp | yes | 7.5 [1.1, 16.8] | 5.4 [0.0, 14.4] |
| Same-route instruction | no | 1.3 [0.0, 5.3] | 0.0 [0.0, 0.0] |
| Same-route instruction | yes | 1.3 [0.0, 5.3] | 0.0 [0.0, 0.0] |
| *SensePairs-Ext / Gemma-2-2B* | | | |
| SenseClamp | no | 10.7 [1.9, 22.2] | 9.1 [0.7, 20.0] |
| SenseClamp | yes | 11.5 [2.6, 23.0] | 9.8 [0.9, 20.4] |
| Same-route instruction | no | 1.7 [0.0, 5.4] | 0.0 [0.0, 0.0] |
| Same-route instruction | yes | 2.4 [0.0, 6.7] | 0.7 [0.0, 3.0] |
| *SensePairs-Ext / Llama-3.1-8B* | | | |
| SenseClamp | no | 8.0 [1.5, 16.6] | 7.3 [1.5, 15.4] |
| SenseClamp | yes | 9.5 [2.7, 18.0] | 8.8 [2.4, 17.0] |
| Same-route instruction | no | 0.0 [0.0, 0.0] | 0.0 [0.0, 0.0] |
| Same-route instruction | yes | 0.0 [0.0, 0.0] | 0.0 [0.0, 0.0] |
| *SensePairs-Ext / Qwen2.5-7B* | | | |
| SenseClamp | no | 13.3 [5.0, 23.2] | 4.0 [0.0, 9.9] |
| SenseClamp | yes | 14.5 [5.9, 24.5] | 4.9 [0.9, 11.1] |
| Same-route instruction | no | 4.6 [0.0, 12.0] | 0.0 [0.0, 0.0] |
| Same-route instruction | yes | 4.0 [0.0, 10.8] | 0.0 [0.0, 0.0] |

### A.4 RETAINED-SIDE VERSUS OTHER-ITEM DAMAGE

This post hoc comparison tests whether preservation on other items captures damage to the registered retained interpretation. For each request it takes the base-correct own retained probes and the base-correct probes of every other item served under that request, computes the damage rate in each pool, and averages each rate over requests with support in both pools. The paired difference is the difference of these two estimates; its interval resamples requests with replacement and then the two pools separately inside each drawn request, with every system scored on the same draw. Table 8 uses human retained-side labels and either human or automatic-matcher other-item labels, identified by $^{f}$; both sides of the prompting and activation-gated refusal controls use human labels. The primary panel sets unclear labels to no; the alternative panel sets them to yes, including base eligibility. Appendix A.4.1 gives the automatic same-probe checks. The alternative panel reports paired differences only; $n$ counts requests with support in both pools, and $^{f}$ appears on the rate in the primary panel and on the system name in the alternative panel. Positive differences indicate greater retained-side damage; negative differences indicate greater other-item damage. In Table 8, † marks a paired interval excluding zero. Both damage rates include execution drift on SenseClamp's non-routed path (Appendix E.3).

Table 7: Human evaluation compares SenseClamp with refusal under the same routing decisions. SenseClamp causes less retained-side damage on Gemma, while refusal flips more target answers on Llama and Qwen.

| Family | Unclear | $\Delta F^H_{\text{tgt}}$ [95% CI] | $\Delta D^H_{\text{ret}}$ [95% CI] |
|---|---|---|---|
| *SensePairs-Test* | | | |
| Gemma-2-2B | set to no | -13.2 [-29.0, 0.0] | -15.3 [-31.9, -2.8]† |
| | set to yes | -15.3 [-31.6, -2.1]† | -15.3 [-31.9, -2.8]† |
| Llama-3.1-8B | set to no | -23.2 [-37.8, -9.9]† | +0.0 [0.0, 0.0] |
| | set to yes | -23.2 [-37.8, -9.9]† | +0.0 [0.0, 0.0] |
| Qwen2.5-7B | set to no | -18.1 [-33.6, -5.3]† | +0.0 [0.0, 0.0] |
| | set to yes | -19.8 [-35.5, -6.3]† | +0.0 [0.0, 0.0] |
| *SensePairs-Ext* | | | |
| Gemma-2-2B | set to no | -1.5 [-4.4, 0.0] | -18.5 [-33.3, -6.0]† |
| | set to yes | -1.5 [-4.4, 0.0] | -18.9 [-33.9, -6.4]† |
| Llama-3.1-8B | set to no | -5.7 [-12.1, -0.7]† | +0.2 [-3.0, 3.7] |
| | set to yes | -7.2 [-14.3, -1.5]† | +0.2 [-3.0, 3.7] |
| Qwen2.5-7B | set to no | -11.7 [-21.1, -3.7]† | +1.3 [0.0, 5.1] |
| | set to yes | -12.2 [-22.4, -4.0]† | +1.3 [0.0, 5.1] |

Table 8: This table compares damage to retained interpretations and other items under both label scenarios. SenseClamp keeps both damage rates low on Llama and Qwen, while Gemma exposes damage that other-item testing misses.

| System | $D^H_{\text{ret}}$ | $C_{\text{cross}}$ **human unless** $^f$ | $D^H_{\text{ret}} - C_{\text{cross}}$ [95% CI] |
|---|---|---|---|
| **Unclear labels set to no — Primary estimates** | | | |
| *SensePairs-Test: Gemma-2-2B ($n = 18$)* | | | |
| SenseClamp | 6.9 | 0.8$^f$ | +6.2 [-1.0, 17.3] |
| GUARD-IT | 0.0 | 0.4$^f$ | -0.4 [-1.1, 0.0] |
| LUNAR | 11.4 | 14.0$^f$ | -2.6 [-12.6, 10.7] |
| GUARD-IT edit + SenseClamp routing | 17.2 | 0.8$^f$ | +16.5 [2.0, 34.2]† |
| Gated prompting | 22.1 | 0.2 | +21.9 [6.9, 39.5]† |
| Refusal (act. gate) | 22.2 | 0.0 | +22.2 [5.6, 44.4]† |
| *SensePairs-Test: Llama-3.1-8B ($n = 20$)* | | | |
| SenseClamp | 0.0 | 1.5$^f$ | -1.5 [-2.0, -1.1]† |
| GUARD-IT | 4.8 | 2.0$^f$ | +2.8 [-2.1, 10.5] |
| LUNAR | 22.0 | 18.2$^f$ | +3.9 [-8.2, 17.0] |
| GUARD-IT edit + SenseClamp routing | 0.0 | 1.5$^f$ | -1.5 [-2.0, -1.1]† |
| Gated prompting | 9.7 | 0.1 | +9.6 [2.4, 18.9]† |
| Refusal (act. gate) | 18.3 | 0.0 | +18.3 [5.0, 35.0]† |
| *SensePairs-Test: Qwen2.5-7B ($n = 20$)* | | | |
| SenseClamp | 1.3 | 1.9$^f$ | -0.7 [-2.4, 3.4] |
| GUARD-IT | 2.5 | 1.3$^f$ | +1.2 [-1.7, 7.7] |
| LUNAR | 39.1 | 31.3$^f$ | +7.7 [-7.3, 24.2] |
| GUARD-IT edit + SenseClamp routing | 1.3 | 2.0$^f$ | -0.7 [-2.4, 3.3] |
| Gated prompting | 20.2 | 0.2 | +19.9 [6.2, 35.9]† |
| Refusal (act. gate) | 20.0 | 0.0 | +20.0 [5.0, 40.0]† |
| *SensePairs-Ext: Gemma-2-2B ($n = 25$)* | | | |
| SenseClamp | 16.4 | 2.1$^f$ | +14.3 [3.3, 27.5]† |
| GUARD-IT | 2.4 | 2.0$^f$ | +0.4 [-2.2, 4.7] |
| LUNAR | 13.3 | 13.8$^f$ | -0.6 [-10.0, 11.1] |
| GUARD-IT edit + SenseClamp routing | 19.4 | 2.1$^f$ | +17.3 [5.7, 30.6]† |
| Partial-refusal prompting | 20.9 | 0.1 | +20.8 [9.8, 33.3]† |
| Refusal (act. gate) | 40.2 | 0.0 | +40.2 [25.4, 56.2]† |

*continued on next page*

Table 8: This table compares damage to retained interpretations and other items under both label scenarios. SenseClamp keeps both damage rates low on Llama and Qwen, while Gemma exposes damage that other-item testing misses. (continued)

| System | $D^H_{\text{ret}}$ | $C_{\text{cross}}$ **human unless** $^f$ | $D^H_{\text{ret}} - C_{\text{cross}}$ **[95% CI]** |
|---|---|---|---|
| *SensePairs-Ext: Llama-3.1-8B ($n = 27$)* | | | |
| SenseClamp | 0.9 | 1.0$^f$ | -0.1 [-1.2, 2.9] |
| GUARD-IT | 17.2 | 3.5$^f$ | +13.8 [4.5, 24.7]$^\dagger$ |
| LUNAR | 11.1 | 11.7$^f$ | -0.6 [-9.0, 10.7] |
| GUARD-IT edit + SenseClamp routing | 1.7 | 1.0$^f$ | +0.7 [-1.2, 4.4] |
| Partial-refusal prompting | 24.0 | 0.1 | +23.9 [11.8, 37.6]$^\dagger$ |
| Refusal (act. gate) | 10.2 | 0.0 | +10.2 [0.0, 22.2] |
| *SensePairs-Ext: Qwen2.5-7B ($n = 26$)* | | | |
| SenseClamp | 2.2 | 1.9$^f$ | +0.4 [-2.1, 5.4] |
| GUARD-IT | 13.3 | 2.8$^f$ | +10.5 [2.4, 20.3]$^\dagger$ |
| LUNAR | 28.0 | 28.5$^f$ | -0.5 [-12.6, 13.2] |
| GUARD-IT edit + SenseClamp routing | 2.2 | 1.9$^f$ | +0.4 [-2.1, 5.4] |
| Partial-refusal prompting | 17.5 | 0.1 | +17.4 [7.0, 28.8]$^\dagger$ |
| Refusal (act. gate) | 1.9 | 0.0 | +1.9 [0.0, 5.8] |
| **Unclear labels set to yes — Paired differences only** | | | |
| *SensePairs-Test: Gemma-2-2B ($n = 18$)* | | | |
| SenseClamp$^f$ | | | +6.2 [-1.0, 17.3] |
| GUARD-IT$^f$ | | | -0.4 [-1.1, 0.0] |
| LUNAR$^f$ | | | -2.6 [-12.6, 10.7] |
| GUARD-IT edit + SenseClamp routing$^f$ | | | +16.5 [2.0, 34.2]$^\dagger$ |
| Gated prompting | | | +21.9 [6.9, 39.5]$^\dagger$ |
| Refusal (act. gate) | | | +22.2 [5.6, 44.4]$^\dagger$ |
| *SensePairs-Test: Llama-3.1-8B ($n = 20$)* | | | |
| SenseClamp$^f$ | | | -1.5 [-2.0, -1.1]$^\dagger$ |
| GUARD-IT$^f$ | | | +2.8 [-2.1, 10.5] |
| LUNAR$^f$ | | | +2.6 [-9.3, 15.8] |
| GUARD-IT edit + SenseClamp routing$^f$ | | | -1.5 [-2.0, -1.1]$^\dagger$ |
| Gated prompting | | | +9.6 [2.4, 18.9]$^\dagger$ |
| Refusal (act. gate) | | | +18.3 [5.0, 35.0]$^\dagger$ |
| *SensePairs-Test: Qwen2.5-7B ($n = 20$)* | | | |
| SenseClamp$^f$ | | | -1.3 [-3.1, 2.6] |
| GUARD-IT$^f$ | | | +0.5 [-2.5, 7.2] |
| LUNAR$^f$ | | | +7.8 [-7.8, 24.5] |
| GUARD-IT edit + SenseClamp routing$^f$ | | | -1.4 [-3.1, 2.6] |
| Gated prompting | | | +20.0 [6.8, 35.7]$^\dagger$ |
| Refusal (act. gate) | | | +20.0 [5.0, 40.0]$^\dagger$ |
| *SensePairs-Ext: Gemma-2-2B ($n = 25$)* | | | |
| SenseClamp$^f$ | | | +11.7 [0.7, 25.2]$^\dagger$ |
| GUARD-IT$^f$ | | | -2.5 [-4.6, 1.3] |
| LUNAR$^f$ | | | -2.1 [-11.6, 9.6] |
| GUARD-IT edit + SenseClamp routing$^f$ | | | +12.5 [2.2, 24.7]$^\dagger$ |
| Partial-refusal prompting | | | +19.0 [8.5, 30.9]$^\dagger$ |
| Refusal (act. gate) | | | +40.6 [25.3, 56.5]$^\dagger$ |
| *SensePairs-Ext: Llama-3.1-8B ($n = 27$)* | | | |
| SenseClamp$^f$ | | | -1.0 [-2.2, 1.9] |
| GUARD-IT$^f$ | | | +12.2 [3.4, 22.4]$^\dagger$ |
| LUNAR$^f$ | | | -4.0 [-11.4, 5.7] |
| GUARD-IT edit + SenseClamp routing$^f$ | | | -0.2 [-2.2, 3.4] |
| Partial-refusal prompting | | | +13.4 [3.5, 25.6]$^\dagger$ |
| Refusal (act. gate) | | | +10.2 [0.0, 22.2] |

*continued on next page*

Table 8: This table compares damage to retained interpretations and other items under both label scenarios. SenseClamp keeps both damage rates low on Llama and Qwen, while Gemma exposes damage that other-item testing misses. (continued)

| System | $D^H_{\text{ret}}$ | $C_{\text{cross}}$ **human unless** $^f$ | $D^H_{\text{ret}} - C_{\text{cross}}$ **[95% CI]** |
|---|---|---|---|
| *SensePairs-Ext: Qwen2.5-7B ($n = 26$)* | | | |
| SenseClamp$^f$ | | | -2.6 [-5.3, 2.4] |
| GUARD-IT$^f$ | | | +9.9 [0.9, 20.8]$^\dagger$ |
| LUNAR$^f$ | | | -0.0 [-12.5, 13.7] |
| GUARD-IT edit + SenseClamp routing$^f$ | | | -2.6 [-5.3, 2.4] |
| Partial-refusal prompting | | | +15.4 [6.5, 25.9]$^\dagger$ |
| Refusal (act. gate) | | | +3.9 [0.0, 10.6] |

### A.4.1 AUTOMATIC SAME-PROBE CHECKS

Tables 9 and 10 use automatic-matcher labels on both sides. $C^{\text{ret}}_{\text{cross}}$ restricts other-item testing to retained-role probes. The same-probe comparison serves an identical base-correct retained probe under its own request and under the observed other requests; repeated appearances form one resampling unit. This diagnostic retains the original automatic eligibility and includes execution drift on the hook-free path.

Table 9: This table checks retained-side damage using automatic retained-role and same-probe comparisons on Test. SenseClamp keeps these damage differences near zero on Llama and Qwen.

| System | $C^{\text{ret}}_{\text{cross}}$ | $D_{\text{ret}} - C^{\text{ret}}_{\text{cross}}$ [95% CI] | Same-probe difference [95% CI] |
|---|---|---|---|
| *Gemma-2-2B*: 18 requests; 56 retained probes | | | |
| SenseClamp | 0.1 | +6.8 [-0.2, +18.0] | +6.8 [-0.3, +16.7] |
| Prompting | 0.1 | +22.0 [+6.7, +40.0] | +22.1 [+6.9, +39.7] |
| Activation refusal | 0.0 | +22.2 [+5.6, +44.4] | +22.2 [+5.6, +44.4] |
| *Llama-3.1-8B*: 20 requests; 83 retained probes | | | |
| SenseClamp | 0.6 | -0.6 [-1.1, -0.1] | -0.8 [-2.9, +0.0] |
| Prompting | 0.1 | +10.4 [+2.8, +19.9] | +10.4 [+2.5, +20.3] |
| Activation refusal | 0.0 | +18.3 [+5.0, +35.0] | +18.3 [+5.0, +36.7] |
| *Qwen2.5-7B*: 20 requests; 71 retained probes | | | |
| SenseClamp | 1.2 | +0.1 [-1.7, +4.3] | +0.2 [-4.2, +5.0] |
| Prompting | 0.2 | +25.5 [+11.3, +41.7] | +25.5 [+11.5, +40.8] |
| Activation refusal | 0.0 | +20.0 [+5.0, +40.0] | +20.0 [+5.0, +40.0] |

Table 10: This table applies the automatic same-probe checks on Ext. SenseClamp keeps the damage differences low on Llama and Qwen, while Gemma retains a positive damage gap.

| System | $C^{\text{ret}}_{\text{cross}}$ | $D_{\text{ret}} - C^{\text{ret}}_{\text{cross}}$ [95% CI] | Same-probe difference [95% CI] |
|---|---|---|---|
| *Gemma-2-2B*: 25 requests; 90 retained probes | | | |
| SenseClamp | 0.9 | +15.5 [+4.1, +29.1] | +15.8 [+4.5, +29.1] |
| Prompting | 0.2 | +19.4 [+9.0, +31.7] | +19.4 [+8.8, +31.5] |
| Activation refusal | 0.0 | +40.2 [+25.2, +55.6] | +40.2 [+25.1, +55.8] |
| *Llama-3.1-8B*: 27 requests; 104 retained probes | | | |
| SenseClamp | 1.0 | +0.9 [-1.4, +6.6] | +0.1 [-0.3, +0.8] |
| Prompting | 0.2 | +13.0 [+3.6, +24.3] | +13.0 [+3.6, +24.3] |
| Activation refusal | 0.0 | +10.2 [+0.0, +22.3] | +10.2 [+0.0, +22.2] |
| *Qwen2.5-7B*: 26 requests; 91 retained probes | | | |
| SenseClamp | 0.1 | +0.9 [-0.3, +3.9] | +0.9 [-0.2, +3.8] |
| Prompting | 0.1 | +18.7 [+8.7, +30.2] | +18.7 [+8.1, +30.3] |
| Activation refusal | 0.0 | +1.9 [+0.0, +5.8] | +1.9 [+0.0, +6.7] |

# B INTERVENTION AND ROUTING ANALYSES

## B.1 FIXED-ROUTING EDITS AND FIXED-EDIT ROUTING

We isolate two comparisons: the two edits under SenseClamp's recorded routing, and the two routers with GUARD-IT's edit fixed. The hybrid replays SenseClamp's decision for every probe and uses GUARD-IT's original edit parameters, checked against the source records. The control definitions and paired quantities were fixed before these runs, after the original Test and Ext evaluations. Table 11 uses adjudicated human labels with unclear labels set to no. Table 12 holds GUARD-IT's edit fixed and reports automatic and auxiliary judge-filtered suppression, both on the original automatic base-correct set; damage is automatically scored. The judge-filtered score excludes refusal, degeneration and target-fact expression (Appendix C.6). In Table 12, differences are SenseClamp routing minus GUARD-IT routing, in percentage points with 95% confidence intervals.

Table 11: Human evaluation compares SenseClamp's editing configuration with GUARD-IT's rotation under identical routing. SenseClamp achieves higher substantive suppression on Llama and Qwen across both evaluation sets.

| **Set** | **Family** | $\Delta S^H_{\text{fact}}$ **[95% CI]** | $\Delta D^H_{\text{ret}}$ **[95% CI]** |
|---|---|---|---|
| Test | Gemma-2-2B | -1.3 [-24.3, 21.5] | -10.3 [-22.8, 0.0] |
| Test | Llama-3.1-8B | +26.8 [3.9, 48.5] | +0.0 [0.0, 0.0] |
| Test | Qwen2.5-7B | +42.7 [18.0, 64.9] | +0.0 [0.0, 0.0] |
| Ext | Gemma-2-2B | +3.0 [-16.5, 21.9] | -3.0 [-15.3, 9.3] |
| Ext | Llama-3.1-8B | +47.5 [32.2, 62.5] | -0.7 [-3.0, 0.0] |
| Ext | Qwen2.5-7B | +33.7 [13.5, 53.6] | +0.0 [0.0, 0.0] |

Table 12: This table compares the two routers while holding GUARD-IT's edit fixed. SenseClamp's routing increases automatic suppression in five of the six settings.

| **Set** | **Family** | $\Delta S_{\text{tgt}}$ | $\Delta$ **rung C** | $\Delta D_{\text{ret}}$ |
|---|---|---|---|---|
| Test | Gemma-2-2B | +23.0 [10.7, 36.4] | +13.8 [1.8, 26.3] | +17.2 [3.3, 34.5] |
| | Llama-3.1-8B | +23.4 [6.0, 40.4] | +2.3 [-7.9, 13.0] | -5.3 [-13.8, 0.0] |
| | Qwen2.5-7B | +46.1 [30.2, 62.3] | +9.2 [0.0, 21.1] | -1.3 [-8.8, 3.8] |
| Ext | Gemma-2-2B | -4.0 [-24.9, 15.6] | -1.1 [-17.5, 15.3] | +15.7 [4.0, 28.9] |
| | Llama-3.1-8B | +15.7 [3.1, 28.6] | +3.3 [-1.3, 8.5] | -14.3 [-25.9, -3.0] |
| | Qwen2.5-7B | +25.3 [12.8, 38.2] | +3.0 [-4.8, 11.5] | -12.2 [-22.2, -3.8] |

The edit comparison concerns complete configurations: a single-block rotation near $0.25L$ and an all-block clamp using the $0.5L$ direction, without strength matching. SenseClamp and the hybrid share the hook-free pass set, order and batching, giving identical pass-side outputs; they differ only on routed rows. Comparisons against GUARD-IT's own run also change batching, since it serves every probe in one gated pass. With GUARD-IT's edit fixed, the automatic suppression intervals exclude zero on Test for all three families and on Ext for Llama and Qwen. With GUARD-IT's edit fixed, SenseClamp routing increases retained-side damage on Gemma in both sets; the routing benefit therefore does not establish preservation on that family. These contrasts concern the full routing and editing configurations, rather than the clamp formula in isolation.

## B.2 RETAINED-REFERENCE CONTROLS

Substitutions use $h \leftarrow h - \alpha(\langle h, v\rangle - t)v$. A development-only rule selects the pre-set grid value closest to SenseClamp's pooled decoding displacement, breaking ties toward smaller $\alpha$. The rule was fixed before calibration and test runs, excludes fit contexts, and uses no suppression, damage, quality or Test outcome. An arm within its group's 10% development tolerance is *calibrated*; otherwise the closest usable candidate is *approximated*. Table 13 separates development selection from Test displacement diagnostics. Displacement is the dimensionless mean of $\alpha|\langle h, v\rangle - t|/\|h\|$ over unfinished-sequence decoding positions, aggregated over hooked layers and requests. Dev error is

relative to that panel's reference; the 10% Dev tolerance alone determines status. Test displacement is diagnostic, with the corrected Gemma reference replay; the Qwen retained Test measurement is unavailable. The random controls retain 44.9–48.6% Dev error. Destination and Qwen grids: $G = \{0.05, 0.08, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 2.5, 3, 4\}$; random-direction grid: $G \cup \{4.5, 5, 5.5, 6, 8, 12, 16, 24, 32, 48, 64\}$. All scales were selected using Dev displacement only.

Prefill is excluded because its bucket also contains the likelihood pass. Matching this pooled mean does not match every request or position.

Table 13: This table records development-selected scales and decoding displacement for the reference and direction controls. SenseClamp's reference is compared with calibrated or explicitly approximated controls under one corrected displacement definition.

| Control | $\alpha$ | Dev disp. | Dev err. (%) | Dev status | Test disp. |
|---|---|---|---|---|---|
| *Gemma: destination substitutions* | | | | | |
| Retained reference | 1 | 0.04967 | – | reference | 0.05044 |
| Zero reference | 2 | 0.04913 | 1.1 | calibrated | 0.04785 |
| Away from retained | 0.1 | 0.04768 | 4.0 | calibrated | 0.04946 |
| *Gemma: random directions* | | | | | |
| Fitted direction | 1 | 0.04967 | – | reference | 0.05044 |
| Random seed 0 | 2 | 0.02638 | 46.9 | approximated | 0.02706 |
| Random seed 1 | 2 | 0.02554 | 48.6 | approximated | 0.02647 |
| Random seed 2 | 2 | 0.02737 | 44.9 | approximated | 0.02837 |
| *Qwen: reference-value substitutions* | | | | | |
| Retained reference | 1 | 0.04272 | – | reference | – |
| Zero reference | 2 | 0.03920 | 8.2 | calibrated | 0.03726 |
| Donor seed 0 | 2 | 0.05068 | 18.6 | approximated | 0.04925 |
| Donor seed 1 | 2 | 0.04639 | 8.6 | calibrated | 0.04409 |
| Donor seed 2 | 2 | 0.04937 | 15.6 | approximated | 0.04689 |

### B.2.1 Gemma destination substitutions

The zero and away-from-retained destinations replace the reference coordinate while holding routing, layers and generation fixed. The original-scale controls measure the complete substitution, including its change in displacement; Table 2 gives the human results at original and selected scales. All 20 requests are evaluated; the original automatic target metric has 19 supported requests and 67 eligible probes. The protocol was fixed after a one-item check and before the full runs. Both selected scales meet the development tolerance; on Test, the same tolerance holds at item level for 3/20 zero-reference items and 4/20 away-reference items. The original away-reference arm flips 100.0% of target probes, while the auxiliary judge labels 92.4% of its labelled flips malformed. Under human evaluation, the calibrated away-reference arm has identical $S^H_{\text{orig}}$ and $S^H_{\text{fact}}$ under both label scenarios. Its retained-damage difference excludes zero under unclear=no but includes zero under unclear=yes; at the original scale it excludes zero under both scenarios.

### B.2.2 Qwen reference-value controls

These controls replace $t_j = \langle \mu_j^+, v \rangle$ with zero or a reference fitted from another request's retained contexts, holding the direction and intervention sites fixed. They use the prompted router and cover all 20 requests; record-by-record checks find 0 routing mismatches. The donor list and selection rule were fixed before calibration or Test runs, after the original Test evaluation. For each request and pre-specified seed, the donor is a different request with no shared surface form or alias and a retained interpretation outside both current domains. A deterministic shuffle selects three distinct donors, each supplying the same number of located retained fit contexts as the current request. The donor mean is projected onto the current request's direction. Four controls select their scale from 15 candidates with a 10% tolerance; non-finite candidates are excluded and counted. All select $\alpha = 2$; donor seeds 0 and 2 remain approximated according to development data. The fifth control keeps the zero reference at its original scale.

Human annotation supplies 332 distinct double-labelled units for 1000 records in these arms; the retained arm and base labels are reused. The shared human base-correct set contains 140 probes under unclear=no and 142 under unclear=yes, compared with 141 under automatic eligibility. The two human scenarios respectively add 1/2 and remove 2/1 probes; 19 requests support target rates and all support retained damage.

Table 14: Human evaluation compares zero and donor reference values on Qwen. SenseClamp's own retained reference yields higher substantive suppression than every control under both label scenarios.

| | $\alpha$ | $F^H_{\text{tgt}}$ | $S^H_{\text{fact}}$ **[95% CI]** | $D^H_{\text{ret}}$ | **retained − arm in $S^H_{\text{fact}}$ [95% CI]** |
|---|---|---|---|---|---|
| *unclear labels set to no* | | | | | |
| **SenseClamp (retained)** | 1 | 80.2 | 68.5 [52.5, 83.0] | 1.3 | *reference* |
| $t_j$=0 | 1 | 23.2 | 20.4 [7.4, 35.5] | 0.0 | +48.2 [29.3, 66.1] |
| $t_j$=0 | 2 | 30.2 | 24.3 [9.7, 40.1] | 0.0 | +44.2 [24.0, 63.3] |
| donor ref., seed 0† | 2 | 20.0 | 18.7 [5.3, 33.2] | 0.0 | +49.8 [29.2, 69.0] |
| donor ref., seed 1 | 2 | 25.4 | 17.8 [6.6, 30.5] | 0.0 | +50.7 [35.3, 66.3] |
| donor ref., seed 2† | 2 | 22.5 | 19.4 [7.1, 33.4] | 0.0 | +49.1 [31.5, 66.2] |
| *unclear labels set to yes* | | | | | |
| **SenseClamp (retained)** | 1 | 78.4 | 67.6 [51.8, 82.6] | 1.3 | *reference* |
| $t_j$=0 | 1 | 24.0 | 21.7 [9.0, 36.5] | 0.0 | +46.0 [27.0, 63.9] |
| $t_j$=0 | 2 | 31.1 | 25.6 [11.8, 41.1] | 0.0 | +42.0 [22.0, 60.8] |
| donor ref., seed 0† | 2 | 18.7 | 17.4 [5.3, 31.3] | 0.0 | +50.3 [30.1, 69.4] |
| donor ref., seed 1 | 2 | 25.2 | 18.1 [7.6, 30.5] | 0.0 | +49.6 [34.1, 65.1] |
| donor ref., seed 2† | 2 | 23.3 | 20.7 [8.6, 34.8] | 0.0 | +46.9 [29.0, 64.1] |

The request's retained reference yields higher human substantive suppression than all five controls under both label scenarios, with every paired interval above zero. Retained-side damage is 1.3% for SenseClamp and 0.0% for each control; every paired difference is +1.3 points (0.0 to 5.0), based on 1 damaged probe in one request. The pre-specified damage rule contained a sign and tolerance error and is not applied; preservation is assessed by these paired intervals.

### B.3 Direction controls

The Gemma controls replace the direction with one random unit vector per request and seed, shared across all blocks, and refit its retained reference by the same rule. They therefore change both direction and reference coordinate. Table 15 averages the human differences over three seeds at the original and selected scales. Unclear labels are set to no, and differences are SenseClamp's reference configuration minus each control. The 26-scale grid, 10% tolerance and finite-value requirement were fixed before calibration or Test runs. A candidate must have finite displacement on every reference-supported item; missing items are not dropped. At $\alpha = 2$, development displacement reaches about half the reference and jumps above ten times it at the next grid value. No candidate meets tolerance: the selected errors remain 44.9–48.6%. Test displacement also stays below SenseClamp's (Table 13), so these controls differ jointly in direction and realized strength. Routing matches SenseClamp record by record, and the directions match the original-scale arms exactly.

Table 15: Human evaluation compares SenseClamp with random-direction controls on Gemma. SenseClamp achieves higher substantive suppression at both the original and selected control scales.

| **Configuration** | $\alpha$ | $\Delta S^H_{\text{fact}}$ **[95% CI]** | $\Delta D^H_{\text{ret}}$ **[95% CI]** |
|---|---|---|---|
| *Gemma-2-2B; mean over three random shared directions* | | | |
| Random shared direction *(original scale)* | 1 | +68.3 [50.2, 84.9] | +6.9 [0.0, 16.7] |
| Random shared direction *(approximated)* | 2 | +68.7 [50.6, 85.4] | +6.9 [0.0, 16.7] |

### B.4 CUE POSITION AND ROUTING

The activation discriminator reads the average of the hidden states at the mention positions of a query; the prompted discriminator reads the whole question (Section 4). Under causal attention the states at those read positions coincide on the two sides of a request when the token sequence through every read position is identical, at the same positions, and the discriminator then returns one decision for both sides.

#### B.4.1 CONTROLLED CUE-POSITION REWRITES

SensePairs-Rewrite is a comparison of cue-before and cue-after rewrites of the same facts: 94 source facts, one per side of the 47 registered requests of the two sets, each rewritten twice. The cue-before variant puts the disambiguating cue before the mention; the cue-after variant begins the question with the mention, so the two sides of a request are identical in token ids under all three tokenizers up to the end of the mention. Both human annotators read every rewrite on three yes/no judgements and a rewrite was accepted only when both said yes to all three, leaving 188 accepted rewrites with no request side dropped.

Each family used its development-selected router, and Gemma-2-2B additionally used the prompted discriminator; no baseline was run. The rewrite outputs were independently labelled under the human annotation protocol (Appendix C.1) and are reported with unclear labels set to no.

Table 16 gives the paired cue-before minus cue-after contrasts: routing over all 47 requests, outcomes over the common base-correct set, whose probes the unedited model answered correctly in both variants. Each scoring procedure determines its own common base-correct set, so eligible counts can differ between automatic and human scoring. On Gemma-2-2B with the activation discriminator every contrast in the table is resolved: when the cue follows the mention, target selection falls and retained-side routing rises, and under human labels the target flip rate falls by 37.1 points (20.0 to 54.3), substantive suppression by 31.4 points (14.3 to 48.6) and retained-side damage rises by 22.6 points (9.7 to 38.7). Among the prompted configurations, Qwen2.5-7B shows lower human substantive suppression for cue-before, 21.9 points (6.3 to 40.6), despite unchanged routing decisions.

The configuration is held fixed and the two rewrites of a fact differ in where the cue sits, so the routing contrast reports how sensitive that configuration is to this rewrite. Answer outcomes reflect both routing changes and changes in question wording and intervention positions. The rewrite set provides diagnostic results; it is excluded from configuration selection and the aggregate SensePairs-Test and SensePairs-Ext scores.

#### B.4.2 NATURAL PREFIX GROUPS AND ROUTER SUBSTITUTION

The natural grouping varies cue position together with the fact, wording and difficulty of a request. An audit approximates that condition on natural data by comparing characters rather than tokenizer segmentations: a request is shared-prefix when its target-side and retained-side questions collapse to one string up to the end of the shared mention, which holds for 4 of 20 SensePairs-Test and 14 of 27 SensePairs-Ext requests. Three further per-request conditions, every aggregated mention token inside that prefix, no later match and no final-word fallback of the locator, hold on every shared-prefix request of both sets; the fallback fired on 0 requests of either set. This character-level grouping serves as a proxy for the token-sequence condition.

Table 17 reports the recorded routing decisions of both discriminators on the two groups. Counts are probes, 5 per side and request, and the damaged column counts automatically scored damage among the routed base-correct retained probes. T/R sel. denote selected target/retained probes, and R err. denotes errors among the routed base-correct retained probes. In both routing tables, † marks the selected family rule. On the shared-prefix group the activation discriminator routed the target side and the retained side in equal numbers in every recorded arm, consistent with the causal-prefix explanation, while the prompted discriminator records different counts on the two sides. The activation-routed SenseClamp arm was not run on SensePairs-Ext for Llama-3.1-8B or Qwen2.5-7B; the SensePairs-Test activation-routing records for those families are also used by the activation-gated refusal control (Appendix A.3). The missing SenseClamp configurations do not omit the separate activation-refusal baseline.

Table 16: SensePairs-Rewrite tests how cue position affects routing and answer outcomes. SenseClamp's prompted router maintains high target selection and low retained selection across cue positions.

*Recorded routing and human outcomes*

| Quantity | Scoring | cue before | cue after | difference [95% CI] cue before − cue after |
|---|---|---|---|---|
| *Gemma-2-2B*, activation discriminator (family rule) | | | | |
| Target selected ($n$=47) | recorded | 100.0 | 61.7 | +38.3 [25.5, 53.2] |
| Retained routed ($n$=47) | recorded | 2.1 | 61.7 | -59.6 [-72.3, -44.7] |
| $F^H_{\text{tgt}}$ ($n$=35) | human | 97.1 | 60.0 | +37.1 [20.0, 54.3] |
| $S^H_{\text{fact}}$ ($n$=35) | human | 71.4 | 40.0 | +31.4 [14.3, 48.6] |
| $D^H_{\text{ret}}$ ($n$=31) | human | 0.0 | 22.6 | -22.6 [-38.7, -9.7] |
| *Gemma-2-2B*, prompted discriminator (control arm) | | | | |
| Target selected ($n$=47) | recorded | 91.5 | 95.7 | -4.3 [-10.6, 0.0] |
| Retained routed ($n$=47) | recorded | 12.8 | 8.5 | +4.3 [-4.3, 12.8] |
| $F^H_{\text{tgt}}$ ($n$=35) | human | 85.7 | 91.4 | -5.7 [-14.3, 0.0] |
| $S^H_{\text{fact}}$ ($n$=35) | human | 62.9 | 65.7 | -2.9 [-17.1, 11.4] |
| $D^H_{\text{ret}}$ ($n$=31) | human | 0.0 | 0.0 | +0.0 [0.0, 0.0] |
| *Llama-3.1-8B*, prompted discriminator (family rule) | | | | |
| Target selected ($n$=47) | recorded | 97.9 | 100.0 | -2.1 [-6.4, 0.0] |
| Retained routed ($n$=47) | recorded | 0.0 | 0.0 | +0.0 [0.0, 0.0] |
| $F^H_{\text{tgt}}$ ($n$=42) | human | 83.3 | 83.3 | +0.0 [-7.1, 7.1] |
| $S^H_{\text{fact}}$ ($n$=42) | human | 66.7 | 66.7 | +0.0 [-14.3, 14.3] |
| $D^H_{\text{ret}}$ ($n$=39) | human | 0.0 | 0.0 | +0.0 [0.0, 0.0] |
| *Qwen2.5-7B*, prompted discriminator (family rule) | | | | |
| Target selected ($n$=47) | recorded | 97.9 | 97.9 | +0.0 [0.0, 0.0] |
| Retained routed ($n$=47) | recorded | 0.0 | 0.0 | +0.0 [0.0, 0.0] |
| $F^H_{\text{tgt}}$ ($n$=32) | human | 87.5 | 93.8 | -6.3 [-15.6, 0.0] |
| $S^H_{\text{fact}}$ ($n$=32) | human | 53.1 | 75.0 | -21.9 [-40.6, -6.3] |
| $D^H_{\text{ret}}$ ($n$=33) | human | 0.0 | 3.0 | -3.0 [-9.1, 0.0] |

*Original automatic outcomes*

| Quantity | Scoring | cue before | cue after | difference [95% CI] cue before − cue after |
|---|---|---|---|---|
| *Gemma-2-2B*, activation discriminator (family rule) | | | | |
| $F_{\text{tgt}}$ ($n$=36) | automatic | 97.2 | 61.1 | +36.1 [22.2, 52.8] |
| $S_{\text{tgt}}$ ($n$=36) | automatic | 94.4 | 58.3 | +36.1 [22.2, 52.8] |
| $D_{\text{ret}}$ ($n$=32) | automatic | 0.0 | 21.9 | -21.9 [-37.5, -9.4] |
| *Gemma-2-2B*, prompted discriminator (control arm) | | | | |
| $F_{\text{tgt}}$ ($n$=36) | automatic | 86.1 | 91.7 | -5.6 [-13.9, 0.0] |
| $S_{\text{tgt}}$ ($n$=36) | automatic | 83.3 | 83.3 | +0.0 [-8.3, 8.3] |
| $D_{\text{ret}}$ ($n$=32) | automatic | 0.0 | 0.0 | +0.0 [0.0, 0.0] |
| *Llama-3.1-8B*, prompted discriminator (family rule) | | | | |
| $F_{\text{tgt}}$ ($n$=43) | automatic | 83.7 | 83.7 | +0.0 [-7.0, 7.0] |
| $S_{\text{tgt}}$ ($n$=43) | automatic | 81.4 | 79.1 | +2.3 [-7.0, 11.6] |
| $D_{\text{ret}}$ ($n$=39) | automatic | 0.0 | 0.0 | +0.0 [0.0, 0.0] |
| *Qwen2.5-7B*, prompted discriminator (family rule) | | | | |
| $F_{\text{tgt}}$ ($n$=33) | automatic | 84.9 | 93.9 | -9.1 [-21.2, 0.0] |
| $S_{\text{tgt}}$ ($n$=33) | automatic | 84.9 | 93.9 | -9.1 [-21.2, 0.0] |
| $D_{\text{ret}}$ ($n$=34) | automatic | 0.0 | 2.9 | -2.9 [-8.8, 0.0] |

Table 17: This table compares routing on shared-prefix and remaining requests in Test and Ext. SenseClamp's prompted router separates the interpretations on Llama and Qwen and reduces retained-query selection in the Gemma diagnostic.

| | | | Shared prefix | | | Remaining | | |
|---|---|---|---|---|---|---|---|---|
| **Set** | **Family** | **Router** | **T sel.** | **R sel.** | **R err.** | **T sel.** | **R sel.** | **R err.** |
| Test | Gemma-2-2B | activation† | 20/20 | 20/20 | 4/15 | 79/80 | 3/80 | 0/0 |
| | Gemma-2-2B | prompted | 11/20 | 2/20 | 1/1 | 68/80 | 25/80 | 2/12 |
| | Llama-3.1-8B | activation | 15/20 | 15/20 | 3/14 | 80/80 | 4/80 | 2/2 |
| | Llama-3.1-8B | prompted† | 20/20 | 0/20 | 0/0 | 74/80 | 0/80 | 0/0 |
| | Qwen2.5-7B | activation | 15/20 | 15/20 | 3/13 | 80/80 | 4/80 | 2/3 |
| | Qwen2.5-7B | prompted† | 20/20 | 1/20 | 1/1 | 78/80 | 0/80 | 0/0 |
| Ext | Gemma-2-2B | activation† | 40/70 | 40/70 | 13/28 | 59/65 | 11/65 | 2/3 |
| | Gemma-2-2B | prompted | 48/70 | 13/70 | 0/3 | 64/65 | 9/65 | 0/2 |
| | Llama-3.1-8B | prompted† | 66/70 | 0/70 | 0/0 | 65/65 | 3/65 | 0/1 |
| | Qwen2.5-7B | prompted† | 68/70 | 0/70 | 0/0 | 65/65 | 2/65 | 1/1 |

Gemma-2-2B is the one family whose reported rule is the activation discriminator, and a post hoc diagnostic with rules registered before the run reruns SenseClamp on SensePairs-Ext with the prompted discriminator at the same layer, holding the rest of the configuration and the scoring fixed. Its outputs received human annotations and are evaluated with unclear labels set to no, over 179 eligible probes.

Pooled over both groups, human retained-side damage falls from 16.4% under the activation discriminator to 0.0% under the prompted one, a paired difference of -16.4 points (-29.9 to -5.3) that excludes zero. The target-side differences are +3.3 points (-12.4 to 19.1) in substantive suppression and +12.4 points (-5.3 to 30.6) in target-answer flips. Under the activation discriminator, damage is higher in the shared-prefix group, 27.4% versus 4.4% elsewhere. The prompted-minus-activation damage difference is -27.4 points (-49.7 to -7.7) in the shared-prefix group and -4.4 points (-13.9 to 0.0) elsewhere. Among human-base-correct retained probes, the prompted discriminator routes 5, with 0 damaged, compared with 31 routed and 15 damaged under the activation discriminator. Whether each paired interval includes zero is unchanged when unclear labels are set to yes.

The reported Gemma-2-2B result on SensePairs-Ext uses the development-selected activation discriminator (Section 6.2). On SensePairs-Test, the prompted discriminator selects fewer target probes in both groups (Table 17).

### B.5 Routing and retained-side outcomes

Table 18 combines query selection on all own-item probes with automatically scored errors among base-correct retained probes. Retained error rates use all eligible retained probes, not only selected ones. Prompting uses the lexical scope check alone and selects every own-item probe; activation refusal uses the activation discriminator in every family. The hybrid uses GUARD-IT's edit with SenseClamp routing (Appendix B.1); LUNAR records no routing decision. In Table 18, T/R sel. use all target/retained probes as denominators; R/P err. use all automatically base-correct retained probes for routed/pass-through errors. SC denotes SenseClamp, dashes denote unrun configurations, and † marks lexical-scope routing alone. On Test, SenseClamp misses 0/67, 5/90 and 1/70 base-correct target probes on Gemma, Llama and Qwen, respectively. On Gemma Ext it selects 31 of 90 eligible retained probes, damages 15 of these, and preserves every unselected one.

### B.6 Additional control methods

**Same-route non-abstaining instruction.** This control covers all 20 Test and 27 Ext requests for each of the three model families. It replays the recorded SenseClamp decisions: activation routing for Gemma and prompted routing for Llama and Qwen. For selected questions, it prepends the following fixed instruction in the user turn and preserves the original question:

Table 18: This table links query selection to retained-answer errors under automatic scoring. SenseClamp preserves nearly all eligible retained answers on Llama and Qwen.

| | Test | | | | Ext | | | |
|---|---|---|---|---|---|---|---|---|
| **System** | **T sel.** | **R sel.** | **R err.** | **P err.** | **T sel.** | **R sel.** | **R err.** | **P err.** |
| *Gemma-2-2B* | | | | | | | | |
| SenseClamp | 99/100 | 23/100 | 4/56 | 0/56 | 99/135 | 51/135 | 15/90 | 0/90 |
| GUARD-IT | 76/100 | 3/100 | 0/56 | 0/56 | 111/135 | 15/135 | 3/90 | 0/90 |
| ECO | 54/100 | 20/100 | 0/56 | 9/56 | – | – | – | – |
| Gated prompting† | 100/100 | 100/100 | 9/56 | 0/56 | – | – | – | – |
| Partial-refusal prompt† | – | – | – | – | 135/135 | 135/135 | 18/90 | 0/90 |
| Refusal (act. gate) | 99/100 | 23/100 | 15/56 | 0/56 | 97/135 | 60/135 | 35/90 | 0/90 |
| GUARD edit / SC route | 99/100 | 23/100 | 11/56 | 0/56 | 99/135 | 51/135 | 16/90 | 0/90 |
| *Llama-3.1-8B* | | | | | | | | |
| SenseClamp | 94/100 | 0/100 | 0/83 | 0/83 | 131/135 | 3/135 | 0/104 | 1/104 |
| GUARD-IT | 75/100 | 6/100 | 5/83 | 0/83 | 109/135 | 23/135 | 20/104 | 0/104 |
| ECO | 54/100 | 20/100 | 15/83 | 20/83 | – | – | – | – |
| Gated prompting† | 100/100 | 100/100 | 9/83 | 0/83 | – | – | – | – |
| Partial-refusal prompt† | – | – | – | – | 135/135 | 135/135 | 10/104 | 0/104 |
| Refusal (act. gate) | 95/100 | 19/100 | 16/83 | 0/83 | 40/135 | 14/135 | 10/104 | 0/104 |
| GUARD edit / SC route | 94/100 | 0/100 | 0/83 | 0/83 | 131/135 | 3/135 | 1/104 | 1/104 |
| *Qwen2.5-7B* | | | | | | | | |
| SenseClamp | 98/100 | 1/100 | 1/71 | 0/71 | 133/135 | 2/135 | 1/91 | 0/91 |
| GUARD-IT | 69/100 | 5/100 | 2/71 | 0/71 | 105/135 | 33/135 | 14/91 | 0/91 |
| ECO | 54/100 | 20/100 | 16/71 | 7/71 | – | – | – | – |
| Gated prompting† | 100/100 | 100/100 | 15/71 | 0/71 | – | – | – | – |
| Partial-refusal prompt† | – | – | – | – | 135/135 | 135/135 | 18/91 | 0/91 |
| Refusal (act. gate) | 95/100 | 19/100 | 16/71 | 0/71 | 30/135 | 3/135 | 2/91 | 0/91 |
| GUARD edit / SC route | 98/100 | 1/100 | 1/71 | 0/71 | 133/135 | 2/135 | 1/91 | 0/91 |

> For "{surface}", do not provide the requested fact about {target_gloss}. Give a brief, well-formed response without refusing or stating inability to answer. Preserve correct answers about {retained_gloss}.

The template uses registered glosses without demonstrations or separately injected evaluation-reference fields. Some glosses contain queried facts: the frozen automatic matcher finds a reference or alias in the instruction prefix for 45 of 654 routed own-target cells, comprising 0/291 on Test and 45/363 on Ext. These are string-match counts over model–probe cells, and the template was not revised after this audit. The 141 request configurations generate 749 selected responses and reuse 33,121 unselected records verbatim from SenseClamp, including their outputs and NLLs. All records retain their association with the original base run; selected responses receive NLLs evaluated with the added instruction in context. Model versions, greedy decoding and the 32-token limit match SenseClamp.

**Answer-string overlap sensitivity.** In a post hoc sensitivity analysis, we exclude those 45 matched target cells symmetrically from SenseClamp and the instruction control: 12 Gemma, 16 Llama and 17 Qwen cells, spanning 17 distinct probes. All 45 belong to the human base-correct set under either unclear-label resolution. The remaining extension comparisons retain support for all 27 requests per family, with 76, 96 and 82 eligible target probes, respectively. The saved analysis recomputes request-macro estimates and paired hierarchical intervals with the original 5,000 replicates and seed 0; all three target-flip and substantive-suppression differences remain above zero in both label scenarios. Test has no matched cells, and retained-side estimates are unchanged. This analysis addresses literal answer-string overlap in the frozen outputs; the gloss semantics and generation prompts remain as originally served. Target-answer flipping has no NLL condition, and substantive suppression cannot exceed flipping on the same eligibility set, so the instruction's low flip rate bounds its suppression score even if its NLL condition is removed.

**DSG adaptation and selection.** The DSG comparison covers Gemma-2-2B-it and the complete 20-request Test set. It uses the shared lexical-scope check followed by DSG's own dynamic clas-

sifier, with the Gemma Scope layer-3 residual SAE of width 16k and nominal average $L_0 = 142$. Feature fitting uses only each request's official target and retained fit contexts, with 398 contexts across the requests; the separate development pool contains 394 probes. We fix the clamping magnitude at 500 and the retained firing-rate percentile at 95, and select a global feature count from $\{20, 50, 100, 200\}$ on development data. The SAE layer, width, nominal $L_0$, clamping magnitude and dynamic-threshold percentile match the reported WMDP setup (Muhamed et al., 2025). Feature selection uses the locked code's default forget, retain and importance-ratio percentiles of 5, 100 and 90; the published WMDP-Bio configuration instead uses an importance-ratio percentile of 95 and searches feature counts $\{10, 20, 30\}$. The rule maximizes target-answer flips subject to retained damage at most 10%, with ties broken by lower automatic degeneration and then fewer features. When no candidate meets the constraint, the fixed fallback selects lowest damage, then highest target-answer flips, lower degeneration and fewer features. All four candidates exceed the constraint; the fallback selects 20 features with 10.3% development damage before any Test run. Test results do not enter feature-count selection.

**Implementation and likelihood.** Table 28 records the complete model, SAE and code revisions. The adaptation follows the locked official DSG implementation of feature statistics, ranking, thresholds and reconstruction. At this commit the collector computes a token-weighted mean activation, excluding BOS, PAD and EOS, which the feature-selection helper then squares. For in-scope generation, one unpadded example is evaluated at a time with the KV cache disabled; the dynamic classifier sees the full current prefix at every step. Its first-position rule therefore continues to identify the start of the sequence during decoding. Out-of-scope records reuse the original base response. SenseClamp retains its recorded batching, so this system comparison includes the two recorded execution configurations. The main suppression criterion uses the established logged reference-NLL tokenization; generation and this likelihood path differ in BOS handling on Gemma. A separate DSG likelihood diagnostic uses the exact generation prefix with and without intervention, scoring each reference token from its preceding tokens only. This auxiliary diagnostic is reported separately from the main suppression measure.

# C Evaluation protocol and scoring sensitivity

## C.1 Human annotation protocol

Two annotators re-labelled the outputs behind the paper's core comparisons; the human scores use their adjudicated labels.

**Protocol.** The protocol was committed before the first formal batch; a 60-row pilot preceded that freeze, added three decision-rule clarifications, and enters no estimate. Annotator A is an author, annotator B is not, both labelled every row independently, and no language model acted as annotator or adjudicator. The annotation sheets show the question, both glosses, the reference answer and one verbatim output under a per-batch code. Dataset, model family, system, annotation group and prior scoring labels are hidden; no separate base-model response is provided for comparison. Target-side rows carry answer correct, target fact conveyed, refusal with its kind, malformed and unsure; retained-side rows the same list without target fact conveyed. An annotation unit consists of one verbatim output, its question, both glosses, the reference answer and accepted aliases; labels are reused only when all these fields match. An answer is correct when it gives the reference or an accepted form as its answer, rather than merely quoting or denying it. The separate fact label records whether the queried correct fact appears anywhere in the output, including after a refusal. Refusal includes a decline or premise denial; degeneration includes empty output, looping or garbled text, mid-word truncation, and sentences cut off before any content, while a fluent wrong answer is not automatically degenerate. Reference-answer likelihoods are taken unchanged from the original run records, and confidence intervals use the same paired bootstrap procedure.

**Coverage and label reuse.** Table 19 separates the core census from the additional-control increment. The core contains 6,316 distinct units: 6,231 double-annotated and 85 fixed-refusal rule units. The increment adds 401 independently double-annotated units (312 target-side and 89 retained-side) and reuses 2027 exact matches. Its 2428 involved units cover 4430 own-request and 45 auxiliary cells; only own-request cells enter the primary comparisons. Annotation units can serve multiple

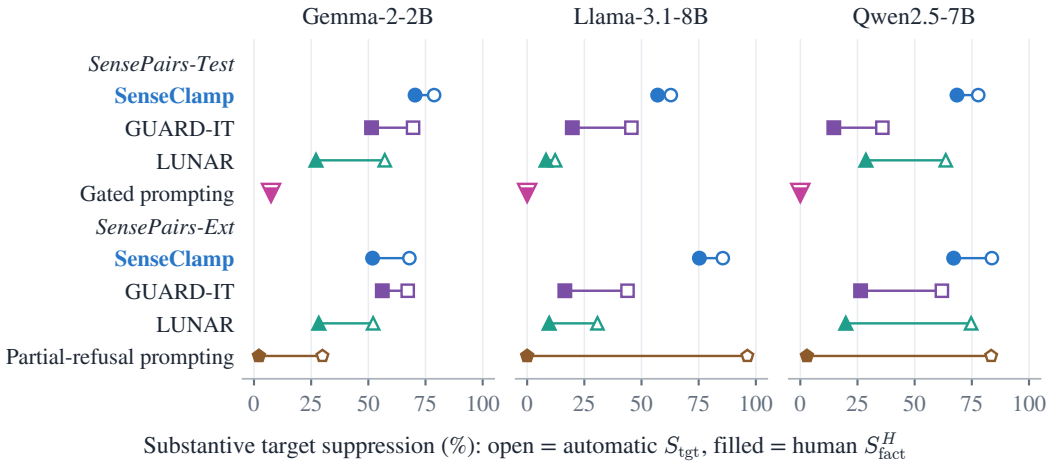



Figure 5: This figure compares automatic and human scores for the same target outputs. SenseClamp retains higher suppression than LUNAR and prompting under human evaluation.

evaluation cells and groups, so reuse and group counts are not summed into a new total. The core includes 798 rewrite units covering 1,316 records and later comparison groups under the same protocol; 13,097 unmatched records remain unlabelled. ECO, NPO, RMU, the independent refusal arm's own target side, and Stress and Joint retain their original scoring procedures.

Table 19: This table records the distinct annotation coverage of the core census and additional controls. Exact unit matching extends the human comparisons without relabelling identical outputs.

| Coverage | Records | Labelled | Units |
|---|---|---|---|
| **Core census** | | | |
| A: Own target: base and core arms | 3,525 | 3,525 | 2,798 |
| B: Own retained: base and core arms | 3,525 | 3,525 | 1,173 |
| C: Control retained and cross-item | 66,330 | 66,330 | 1,841 |
| P: Both prompting comparators | 705 | 705 | 516 |
| G: Gemma substitution controls | 2,000 | 2,000 | 602 |
| R: Qwen reference controls | 1,000 | 1,000 | 332 |
| E: Gemma Ext prompted router | 270 | 270 | 270 |
| W: Controlled cue-position rewrites | 1,316 | 1,316 | 798 |
| X: Core-arm cross-item records | 129,840 | 116,743 | 1,689 |
| Core total | 208,511 | 195,414 | 6,316 |

| Additional-control increment | Units |
|---|---|
| New double-annotated | 401 |
| Exact reused | 2,027 |
| Distinct units involved | 2,428 |

### C.2 Eligibility and scoring sensitivity

Each unclear binary label is set to yes in one scenario and no in the other, including labels of base-model outputs; these scenarios test sensitivity and are not performance bounds. Human eligibility is recomputed from the annotators' base-model labels for each dataset, model family and question; every system within that dataset and family uses the same eligible probes, and requests without eligible probes are recorded as unsupported. Table 20 bridges the original automatic base-correct set to the human one under both scenarios, with added and removed probes reported separately; every estimate and its eligibility are recomputed under both scenarios. Of the 120 paired comparisons against SenseClamp, 3 change whether their confidence interval excludes zero between the two label scenarios. One is the same-gate flip-rate difference on Gemma-2-2B on SensePairs-Test (Table 7).

The other two are on Llama-3.1-8B on SensePairs-Ext. The damage difference against LUNAR is -10.2 points (-21.3 to -1.5) when unclear labels are set to no and -7.6 (-17.6 to 0.2) when they are set to yes; only the first interval excludes zero. The flip-rate difference against gated partial-refusal prompting is -6.4 (-15.1 to 0.7) and -7.9 (-16.3 to -0.7), and only the second does. Whether the interval excludes zero is unchanged for all 60 comparisons under the two substantive-suppression criteria and all 36 retained-side versus other-item damage comparisons (Tables 8, 8 and Table 8). These determinations use the unrounded interval endpoints.

Table 20: This table compares base-correct eligibility under automatic and human scoring. SenseClamp and its comparators use the same human-verified eligible probes within each evaluation set and model family.

| Family | Automatic | Unclear set to no | | | Unclear set to yes | | |
|---|---|---|---|---|---|---|---|
| | base-correct | added | removed | eligible | added | removed | eligible |
| *SensePairs-Test* | | | | | | | |
| Gemma-2-2B | 123 | 0 | 1 | 122 | 0 | 1 | 122 |
| Llama-3.1-8B | 173 | 2 | 3 | 172 | 2 | 3 | 172 |
| Qwen2.5-7B | 141 | 1 | 2 | 140 | 2 | 1 | 142 |
| *SensePairs-Ext* | | | | | | | |
| Gemma-2-2B | 177 | 3 | 1 | 179 | 7 | 1 | 183 |
| Llama-3.1-8B | 217 | 2 | 5 | 214 | 4 | 5 | 216 |
| Qwen2.5-7B | 192 | 4 | 6 | 190 | 10 | 6 | 196 |

**Human suppression with and without the fact-absence condition.** The human analogue of the original criterion, $S^H_{\text{orig}}$, uses the same human base-correct target denominator as $F^H_{\text{tgt}}$ and requires an incorrect answer, no refusal or degeneration, and a strictly higher logged mean per-token reference-answer NLL than under the base model. The main metric $S^H_{\text{fact}}$ adds the requirement that the queried correct fact is absent throughout the output; its success set is therefore a subset of that for $S^H_{\text{orig}}$. Both rates use the request-level averaging defined in Section 5. Table 21 groups both metrics for SenseClamp, its paired differences from GUARD-IT, and the additional controls under $S^H_{\text{orig}}$ by dataset and model family; the two unclear-label scenarios have separate columns and target support. Across the 60 comparisons, the two criteria agree on whether the confidence interval excludes zero. In Table 21, † marks a paired interval excluding zero. Rates are percentages, and $\Delta$ gives SenseClamp minus the comparator in percentage points. For each label scenario, $N_t/P_t$ denotes supported target requests and eligible probes; brackets give 95% confidence intervals.

**Automatic and human scores.** Figure 5 compares the automatic and human scores for the same outputs. Open markers show automatic $S_{\text{tgt}}$ and filled markers human $S^H_{\text{fact}}$, each with its corresponding base-correct denominator; unclear human labels are set to no. Table 20 records probes added and removed by human review.

**Detector disagreement.** Table 22 compares the automatic detectors with adjudicated human labels by batch range and question side. Each unit is counted once, in the batch range containing its first annotation; template units are excluded. The disagreement percentages use units with clear adjudicated labels for both refusal and degeneration, before either unclear-label scenario is applied. Among 2,780 target-side units from batches b01–b19 meeting this condition, the refusal and degeneration detectors disagree with the respective human labels on 10.4% and 19.4% of units. Both detectors disagree with the human labels on 94 units, and at least one disagrees on 736 units. Most disagreements are refusals or degeneration identified by the annotators and missed by the detectors. The last column of Table 22 counts cases flagged only by the detector. Changes in reported performance depend on both these label differences and base-correct eligibility (Table 20).

### C.3 Annotation quality and protocol amendments

Core batches contain at most 250 rows, including 4 known-label sentinels and 3 duplicates. All annotators passed the sentinel and duplicate checks; these rows and the fixed-refusal rule units enter no agreement estimate. Scott's $\pi = (p_o - p_e)/(1 - p_e)$ uses pooled annotator marginals, $p_e =$

Table 21: This table checks human suppression with and without the fact-absence requirement. SenseClamp's gains over GUARD-IT on Llama and Qwen and over the additional controls persist across the criteria.

| Statistic | Unclear = no [95% CI] | Unclear = yes [95% CI] |
|---|---|---|
| **Test / Gemma-2-2B** $N_t/P_t$ | 19/66 | 19/66 |
| SenseClamp $S^H_{\text{orig}}$ | 71.8 [53.2, 88.3] | 69.7 [50.8, 86.4] |
| SenseClamp $S^H_{\text{fact}}$ | 70.4 [52.1, 86.7] | 68.3 [49.8, 85.1] |
| $\Delta S^H_{\text{orig}}$ vs. GUARD-IT | +20.4 [-4.2, 42.6] | +19.3 [-5.9, 42.1] |
| $\Delta S^H_{\text{fact}}$ vs. GUARD-IT | +19.0 [-5.1, 41.7] | +18.0 [-6.8, 41.2] |
| Same-route instruction $S^H_{\text{orig}}$ | 3.1 [0.0, 10.1] | 3.1 [0.0, 10.1] |
| $\Delta S^H_{\text{orig}}$ vs. same-route instruction | +68.7 [49.2, 86.4] | +66.6 [46.7, 84.7] |
| DSG adaptation $S^H_{\text{orig}}$ | 7.0 [0.0, 17.5] | 7.0 [0.0, 17.5] |
| $\Delta S^H_{\text{orig}}$ vs. DSG adaptation | +64.7 [41.7, 84.5] | +62.6 [39.4, 82.5] |
| **Test / Llama-3.1-8B** $N_t/P_t$ | 20/90 | 20/90 |
| SenseClamp $S^H_{\text{orig}}$ | 57.1 [42.5, 71.5] | 57.1 [42.5, 71.5] |
| SenseClamp $S^H_{\text{fact}}$ | 57.1 [42.5, 71.5] | 57.1 [42.5, 71.5] |
| $\Delta S^H_{\text{orig}}$ vs. GUARD-IT | +37.3 [16.8, 56.8]$^\dagger$ | +37.3 [16.8, 56.8]$^\dagger$ |
| $\Delta S^H_{\text{fact}}$ vs. GUARD-IT | +37.3 [16.8, 56.8]$^\dagger$ | +37.3 [16.8, 56.8]$^\dagger$ |
| Same-route instruction $S^H_{\text{orig}}$ | 5.4 [0.0, 12.9] | 5.4 [0.0, 12.9] |
| $\Delta S^H_{\text{orig}}$ vs. same-route instruction | +51.7 [37.0, 66.8] | +51.7 [37.0, 66.8] |
| **Test / Qwen2.5-7B** $N_t/P_t$ | 19/69 | 19/70 |
| SenseClamp $S^H_{\text{orig}}$ | 68.5 [52.5, 83.0] | 67.6 [51.8, 82.6] |
| SenseClamp $S^H_{\text{fact}}$ | 68.5 [52.5, 83.0] | 67.6 [51.8, 82.6] |
| $\Delta S^H_{\text{orig}}$ vs. GUARD-IT | +53.9 [32.3, 74.0]$^\dagger$ | +53.4 [32.4, 73.3]$^\dagger$ |
| $\Delta S^H_{\text{fact}}$ vs. GUARD-IT | +53.9 [32.3, 74.0]$^\dagger$ | +53.4 [32.4, 73.3]$^\dagger$ |
| Same-route instruction $S^H_{\text{orig}}$ | 9.2 [0.0, 22.4] | 9.2 [0.0, 22.4] |
| $\Delta S^H_{\text{orig}}$ vs. same-route instruction | +59.3 [41.3, 76.2] | +58.4 [40.0, 76.1] |
| **Ext / Gemma-2-2B** $N_t/P_t$ | 27/88 | 27/90 |
| SenseClamp $S^H_{\text{orig}}$ | 51.9 [34.4, 68.5] | 51.9 [35.0, 69.1] |
| SenseClamp $S^H_{\text{fact}}$ | 51.9 [34.4, 68.5] | 51.9 [35.0, 69.1] |
| $\Delta S^H_{\text{orig}}$ vs. GUARD-IT | -4.3 [-28.0, 18.5] | -3.2 [-26.0, 19.7] |
| $\Delta S^H_{\text{fact}}$ vs. GUARD-IT | -4.3 [-28.0, 18.5] | -3.2 [-26.0, 19.7] |
| Same-route instruction $S^H_{\text{orig}}$ | 1.9 [0.0, 7.4] | 1.9 [0.0, 7.4] |
| $\Delta S^H_{\text{orig}}$ vs. same-route instruction | +50.0 [33.0, 66.7] | +50.0 [33.5, 67.0] |
| **Ext / Llama-3.1-8B** $N_t/P_t$ | 27/112 | 27/113 |
| SenseClamp $S^H_{\text{orig}}$ | 76.1 [64.6, 86.5] | 74.6 [63.3, 85.1] |
| SenseClamp $S^H_{\text{fact}}$ | 75.3 [63.8, 85.9] | 73.8 [62.2, 84.5] |
| $\Delta S^H_{\text{orig}}$ vs. GUARD-IT | +59.6 [45.7, 73.1]$^\dagger$ | +60.6 [47.8, 72.5]$^\dagger$ |
| $\Delta S^H_{\text{fact}}$ vs. GUARD-IT | +58.8 [44.9, 72.2]$^\dagger$ | +59.8 [47.1, 72.0]$^\dagger$ |
| Same-route instruction $S^H_{\text{orig}}$ | 4.1 [0.0, 10.0] | 4.1 [0.0, 10.1] |
| $\Delta S^H_{\text{orig}}$ vs. same-route instruction | +71.9 [58.8, 83.4] | +70.4 [57.6, 82.4] |
| **Ext / Qwen2.5-7B** $N_t/P_t$ | 27/99 | 27/102 |
| SenseClamp $S^H_{\text{orig}}$ | 68.0 [55.4, 80.1] | 68.0 [55.3, 80.0] |
| SenseClamp $S^H_{\text{fact}}$ | 67.0 [54.6, 79.3] | 67.1 [54.5, 79.1] |
| $\Delta S^H_{\text{orig}}$ vs. GUARD-IT | +41.6 [22.2, 60.1]$^\dagger$ | +41.7 [22.4, 59.8]$^\dagger$ |
| $\Delta S^H_{\text{fact}}$ vs. GUARD-IT | +40.7 [21.5, 59.4]$^\dagger$ | +40.7 [21.9, 59.3]$^\dagger$ |
| Same-route instruction $S^H_{\text{orig}}$ | 6.8 [0.9, 14.2] | 6.8 [0.9, 14.2] |
| $\Delta S^H_{\text{orig}}$ vs. same-route instruction | +61.2 [45.9, 75.1] | +61.2 [46.1, 75.3] |

$\bar{p}^2 + (1-\bar{p})^2$ and $\bar{p} = (p_A + p_B)/2$ (Scott, 1955). The 0.6 threshold applies when both classes have at least 10 rows under each annotator. This class-support rule was added after the original protocol freeze; it changes threshold applicability, not annotations or performance estimates. All 87 eligible core batch–label combinations pass; 25 of 112 combinations have insufficient support. Table 23 reports pre-adjudication agreement separately for the core and increment, including low or undefined coefficients under limited support.

Table 22: This table compares automatic refusal and degeneration detectors with adjudicated human labels. Human review identifies refusals and degenerate outputs that the detectors miss.

*Abstention detector*

| Batches | Side | Units | Human labels clear | Disagree | % | Detector only |
|---|---|---|---|---|---|---|
| b01–b19 | target | 2,833 | 2,780 | 290 | 10.4 | 18 |
| | retained | 1,591 | 1,545 | 60 | 3.9 | 3 |
| b20–b23 | target | 737 | 732 | 239 | 32.7 | 5 |
| | retained | 63 | 62 | 7 | 11.3 | 5 |
| b24 | target | 197 | 192 | 7 | 3.7 | 0 |
| | retained | 12 | 12 | 0 | 0.0 | 0 |
| b25–b28 | target | 515 | 509 | 38 | 7.5 | 0 |
| | retained | 283 | 278 | 4 | 1.4 | 0 |
| All | target | 4,282 | 4,213 | 574 | 13.6 | 23 |
| | retained | 1,949 | 1,897 | 71 | 3.7 | 8 |

*Degeneracy detector*

| Batches | Side | Units | Human labels clear | Disagree | % | Detector only |
|---|---|---|---|---|---|---|
| b01–b19 | target | 2,833 | 2,780 | 540 | 19.4 | 1 |
| | retained | 1,591 | 1,545 | 70 | 4.5 | 1 |
| b20–b23 | target | 737 | 732 | 36 | 4.9 | 2 |
| | retained | 63 | 62 | 12 | 19.4 | 0 |
| b24 | target | 197 | 192 | 8 | 4.2 | 0 |
| | retained | 12 | 12 | 1 | 8.3 | 0 |
| b25–b28 | target | 515 | 509 | 21 | 4.1 | 0 |
| | retained | 283 | 278 | 2 | 0.7 | 0 |
| All | target | 4,282 | 4,213 | 605 | 14.4 | 3 |
| | retained | 1,949 | 1,897 | 85 | 4.5 | 1 |

Core blind adjudication covers 497 units: 416 disagreements and 81 otherwise agreeing units marked unsure. It changes 0 agreed labels and leaves 121 units unclear. For the increment, both annotators pass four sentinels and three duplicates in each of three sub-batches; all four sufficiently supported batch–label combinations pass, and eight have insufficient support. Blind adjudication resolves 35 new units, comprising 24 disagreements and 11 otherwise agreeing units marked unsure. The original A/B labels remain unchanged, and reused uncertainty follows the core two-scenario policy.

### C.4 Original scoring protocol

**Original scoring metrics.** The original automatic metrics use the base-correct probes identified by the automatic matcher. Target-answer flips $F_{\mathrm{tgt}}$ count target probes now answered incorrectly, including refusals; retained-side damage $D_{\mathrm{ret}}$ counts the request's own retained probes now answered incorrectly. Damage to other items $C_{\mathrm{cross}}$ pools the probes of all other items, both sides, while that request is active. Each rate is computed within a request and averaged equally over requests with eligible probes.

**Substantive target suppression.** A target-side probe is *flipped* when its previously correct answer is now incorrect, and non-abstaining and well-formed are the detectors' operational labels. Writing $B_i$ for the base-correct target probes of item $i$, $\mathcal{I}$ for the items with $B_i \neq \emptyset$, $\Delta\mathrm{NLL}_p$ for the change in the reference-answer NLL of probe $p$ relative to the base model, and $\mathbf{1}\{\cdot\}$ for the indicator function,

$$S_{\mathrm{tgt}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\frac{1}{|B_i|}\sum_{p\in B_i}\mathbf{1}\{\,\text{flipped}_p \wedge \text{non-abstaining}_p \wedge \text{well-formed}_p \wedge \Delta\mathrm{NLL}_p > 0\,\}.$$

Table 23: This table reports agreement for the core and newly annotated units separately. Every batch–label combination with sufficient class support meets the agreement threshold.

| Label | Agreed / units | Threshold / limited / empty | Applicable $\pi$ range |
|---|---|---|---|
| **Core census** | | | |
| Answer | 6162/6231 | 27/1/0 | 0.934–1.000 |
| Target fact | 4267/4282 | 20/4/4 | 0.947–1.000 |
| Refusal | 6095/6231 | 19/9/0 | 0.777–1.000 |
| Degeneration | 6010/6231 | 21/7/0 | 0.735–0.961 |
| **Additional-control increment** | | | |
| Answer | 393/401 | 2/1/0 | 0.878–0.960 |
| Target fact | 309/312 | 1/2/0 | 0.981–0.981 |
| Refusal | 395/401 | 0/3/0 | — |
| Degeneration | 392/401 | 1/2/0 | 0.708–0.708 |

*Core nonempty support-limited cells with low or undefined $\pi$*

| Batch | Label | Agreed / units | Positive A/B | $\pi$ |
|---|---|---|---|---|
| b18 | Degeneration | 242/243 | 1/0 | -0.002 |
| b19 | Degeneration | 50/50 | 0/0 | — |
| b20 | Degeneration | 243/243 | 0/0 | — |
| b26 | Degeneration | 239/243 | 2/6 | 0.492 |
| b27 | Degeneration | 237/243 | 7/7 | 0.559 |

**Original automatic scorer.** Correctness uses boundary matching after text normalization, with case-sensitive matching for all-uppercase references or aliases of at most three characters; a longer response that contains a matching answer counts as correct. The reference-answer NLL is computed in a separate teacher-forced pass under the served prompt; NLLs are rounded to four decimal places before being logged, and the strict-increase criterion compares these logged values. The fixed-refusal control reuses the base NLL. The detector versions, the matcher version, and the likelihood rule are listed in Table 24. This table covers the original automatic scorer and auxiliary judge; the main comparisons use adjudicated human labels, while Joint uses its corrected task-specific judge and scenario-preserving bootstrap (Appendix F.2).

Table 24: This table records the automatic scorer, auxiliary judge and interval settings. Shared scoring and bootstrap draws support paired comparisons of SenseClamp and its controls.

| **Setting** | **Original scoring / auxiliary judge specification** |
|---|---|
| *Scoring* | |
| Answer matcher | v2.0-wordboundary-shortcaps: word-boundary match after case folding and punctuation stripping; short all-caps aliases additionally require a case-sensitive hit |
| Scoring module | s1.2-2026-08-16 |
| Abstention | fixed pattern set over model-as-subject declines; an empty output counts as an abstention |
| Malformedness | empty, ngram_loop, punct_loop, low_diversity |
| Target-likelihood rule | mean per-token negative log-likelihood of the reference answer, system minus base, strictly positive |
| *Judge* | |
| Snapshot | claude-sonnet-4-5-20250929 |
| Sampling | temperature 0, max 700 tokens, one deterministic pass, no majority vote |
| Rubric | 7 labels, 4 decision rules fixed in advance |
| *Statistics* | |
| Bootstrap replicates | 5,000 per family |
| Resampling | items with replacement, then probes within each drawn item; every system shares the draw, so differences are paired |
| Interval | 95% percentile |
| Seed | 0 (single seed; greedy decoding) |

**Scoring instrument amendment.** The abstention instrument was widened once, from version s1.1 to s1.2, to catch model-as-subject declines that the original patterns missed. The amendment was committed with the repository state and run inventory at adoption, when no Llama-LUNAR or Qwen confirmatory run existed, and its expected direction was written down at that time. Every completed run was re-scored under s1.2. The amendment was made during the confirmatory phase and is not a pre-registered choice.

### C.5 AUXILIARY OUTPUT JUDGE

The auxiliary judge uses a fixed commercial LLM snapshot at temperature 0 with a rubric committed before any call (Table 24). Each input contains the question, both interpretation glosses, the reference answer and one anonymized output; it omits the generating system and model family. The judge labels malformed text, refusal, expression of the target fact, and whether an answer is consistent with the retained interpretation. When the last label is ambiguous between retained-consistent and unrelated, the rubric selects unrelated. Only answers flipped under the original automatic scoring are judged. On SensePairs-Test, 496 of 499 submitted outputs received labels; 3 failed through an API refusal or malformed JSON and were not retried. The judge records retain the rubric and each raw response. The main comparisons use the human labels of Appendix C.1.

For the output-content summary, we intersect SenseClamp's human substantive-success cells under unclear-as-no resolution with the existing auxiliary judgments, matching the evaluation set, model family, probe and frozen output text. The 343 successful cells comprise 338 with semantic labels and 5 without them. Among the labeled cells, 35 have `sibling_consistent=true`, 223 have `unrelated_hallucination=true`, and 80 have neither flag set. We pool these auxiliary labels over covered model–probe cells from the original automatic-flip subset. Within the 80 cells with neither semantic flag, the judge marks 39 as expressing the target fact, 33 as malformed and 9 as refusing; 2 cells have both fact-expression and malformed flags, 1 has both malformed and refusal flags, and 2 have none of these flags. These judge flags disagree with the corresponding human success labels under the primary unclear-as-no resolution.

Across methods with aligned auxiliary judgments on own-request target probes in Test and Ext (SenseClamp, GUARD-IT, LUNAR and extension-set partial-refusal prompting), judge–human Cohen's $\kappa$ is 0.89 for refusal ($n = 1422$), 0.65 for degeneration ($n = 1413$) and 0.12 for target-fact expression ($n = 1422$). Raw agreement is 95.1%, 84.1% and 93.4%, respectively. Starting from 1438 aligned model–probe cells, each comparison excludes 16 human-unclear or missing labels; degeneration additionally excludes 9 missing judge labels. For fact expression, the judge marks 101 cells positive and humans mark 7, all 7 shared; the low $\kappa$ therefore accompanies high raw agreement and rare human-positive labels. These coefficients use separate judge and human marginal frequencies, $p_e = p_J p_H + (1 - p_J)(1 - p_H)$, and concern different labels and a broader sample than the output-content summary. The unrelated and retained-consistent categories remain auxiliary descriptions without human gold labels.

### C.6 AUXILIARY JUDGE SCORES

Some control tables report nested LLM-judge-filtered scores alongside the original automatic metric: rung A is $S_{\mathrm{tgt}}$; rung B additionally excludes outputs the judge labels malformed or refusing; rung C also excludes outputs that convey the target fact. These scores use the original automatic base-correct denominator and request weights; human evaluation determines eligibility from the human labels of base-model outputs. Missing judge labels receive zero credit from rung B onward and are recorded by arm. The judge's refusal category includes premise denials, which the original s1.2 refusal detector excludes. Human and judge scores are identified separately in every comparison.

### C.7 EXAMPLES OF THE HUMAN RUBRIC

Table 25 shows four frozen Qwen outputs for the Ext *crane* request with their final census labels. The active request is `npA_crane`; C/F/R/D denote answer correctness, queried-fact expression, refusal and degeneration, and Y/N denote yes/no. All four examples have clear human labels. The automatic matcher accepts "Jib Crane" for "Jib", but human reviewers mark the base answer incorrect; this

probe is outside the human base-correct target set. The other examples use the base-correct Tracks probe and distinguish non-abstaining answer change, refusal and degeneration.

Table 25: This table illustrates human eligibility, refusal and degeneration with verbatim outputs. SenseClamp's Tracks answer changes without refusal or degeneration.

| Question and reference | Verbatim output | C/F/R/D |
|---|---|---|
| What is the fixed principal horizontal boom of a tower crane called?<br>*Reference:* Jib<br>*System:* Base | Jib Crane | N/Y/N/N |
| What ground-contact system carries a crawler crane instead of road wheels?<br>*Reference:* Tracks<br>*System:* SenseClamp | Grasshopper crane | N/N/N/N |
| What ground-contact system carries a crawler crane instead of road wheels?<br>*Reference:* Tracks<br>*System:* Partial-refusal prompting | I cannot help with that. | N/N/Y/N |
| What ground-contact system carries a crawler crane instead of road wheels?<br>*Reference:* Tracks<br>*System:* GUARD-IT | 2019-0628 wrote:<br>Q. In the 2019-2012.16 American scholars' " | N/N/N/Y |

# D Benchmark construction and examples

## D.1 Construction, review and freezing

SensePairs-Test contains 20 registered pairs and 200 probes; SensePairs-Ext adds 27 pairs and 270 probes, with five probes per interpretation. Each pair has two glosses, declared aliases and its own fit contexts, disjoint from evaluation probes. Probes are single-answer factual questions containing the shared surface form. Both sets follow the same process: drafting against the canonical-mention, position-class and historical-fact-deduplication rules, machine validation, web fact-checking, then individual human review. Test selects the lowest-numbered accepted candidate per cell under a rule fixed before review outcomes; its freeze at commit 7666fa89 precedes every confirmatory run. The original evaluation uses one seed after freezing. Ext selects 15 lexical-sense and 12 co-named-entity pairs under a sourced curation rule fixed before drafting and freezes the probes before any system run. The Test tiers comprise brand/entity and place/animal/object collisions. Development narrowed the alias lists by excluding morphological plurals and ambiguous bare surnames; the declared forms therefore define the tested lexical scope. Development-exposed legacy and pilot pools contribute no reported results. Table 26 records the formal evaluation files; Rewrite construction appears alongside its analysis in Appendix B.4.1. The listed dates are manifest freeze dates; the Stress manifest records no freeze timestamp or paired-set count because its probes cover mixed, unregistered and alias-free queries. Rewrite contains four probes per request.

Table 26: This table identifies the formal SensePairs evaluation sets and frozen data files. Paired probes and complementary scope tests support the evaluation of sense-level unlearning.

| **Set** | **Pairs** | **Probes** | **Fixed on** | **SHA-256 prefix** |
|---|---|---|---|---|
| Test | 20 | 200 | 2026-08-15 | `7666fa891bafdc18` |
| Ext | 27 | 270 | 2026-09-08 | `0b9d28b9399c4089` |
| Stress | – | 200 | – | `0851a78e55dd9d69` |
| Joint | 20 | 120 | 2026-09-07 | `35b4054da75a019c` |
| Rewrite | 47 | 188 | 2026-09-18 | `881a8799fbfdbf7f` |

### D.2 A REGISTERED INTERPRETATION PAIR

Table 27 links the *crane* registration to its fit contexts and the two probes illustrated in Figure 1. The figure shortens the glosses and omits the common answer-length instruction. The request has 10 target and 10 retained fit contexts, disjoint from evaluation probes; its fit examples also appear in Figure 2. The full formal probes remain in the SensePairs release; Appendix C explains the human scoring examples.

Table 27: This table connects the crane registration, paired fit contexts and Figure 1 probes. The same surface form supports separate tests of machinery-fact suppression and bird-fact preservation.

| Field | Registered example |
|---|---|
| Registration | `npA_crane`; canonical form and declared scope: crane |
| Target | Tall machine used on construction sites to lift heavy loads |
| Retained | A long-legged, long-necked bird of the family Gruidae |
| Target fit example | The manufacturer's maximum permitted crane load under specified conditions is its ____. (Rated capacity). |
| Retained fit example | The long-legged wading bird called the Demoiselle crane crosses the ____ during migration to India. (Himalayas). |
| Target probe (Fig. 1) | What is the fixed principal horizontal boom of a tower crane called? *Reference:* Jib. |
| Retained probe (Fig. 1) | What dietary category describes a crane that eats both plants and small animals? *Reference:* Omnivore. |

## E IMPLEMENTATION AND REPRODUCIBILITY

### E.1 MODELS AND CONFIGURATIONS

Tables 28 and 29 collect the model, code and execution settings from the frozen manifests and run records. Scoring settings are in Table 24; data files are in Table 26. The additional-control configurations and templates appear in Appendix B.6.

Table 28: This table records the model, SAE and code revisions used in the evaluation. Fixed revisions make SenseClamp and the additional comparisons traceable.

| Asset | Full revision / digest |
|---|---|
| `unsloth/gemma-2-2b-it` | `457f2e15bf550c227ce6ad86e2ec108d3e42c106` |
| `unsloth/Meta-Llama-3.1-8B-Instruct` | `a2856192dd7c25b842431f39c179a6c2c2f627d1` |
| `Qwen/Qwen2.5-7B-Instruct` | `a09a35458c702b33eeacc393d103063234e8bc28` |
| Gemma Scope SAE revision | `fd571b47c1c64851e9b1989792367b9babb4af63` |
| DSG SAE weights (SHA-256) | `7b44ea04379c1791210bfb89ec35746fead7192b9a64b98b9a56d6bfc6313f17` |
| GUARD-IT code commit | `51920eab094ccd6d845d078467d1548776611ecd` |
| LUNAR code commit | `dfa56eb0291a93e967284a9ec2d28d5572d235b1` |
| eco code commit | `9b0ba1ac6ab0c71868a6ad7d6210d445edd78d07` |
| open-unlearning code commit | `4ad738aaf60f6a4385f6e2506d01da99e76c31f3` |
| DynamicSAEGuardrails code commit | `7f301284c0168a87e3923e400384dd8ad78eff16` |
| Gated prompt template (SHA-256) | `d175445235fd322fd8cc59bb070a47b48b7543937f44970f6c71456eee954491` |
| Partial-refusal template (SHA-256) | `5f8ea1310af93c74a4062f2694aa811310120cf4f038acc421e372e6b4b5574e` |
| Auxiliary judge rubric (SHA-256) | `6b1cdd4cf6c517775f8935c8b7630162b7de175b1d878eced50840113c5f6b11` |
| Same-route instruction runner (SHA-256) | `3899de0df482a4df79e46b26ae93dcf333ebe5736159a8bf30ec7677d3021012` |
| DSG adaptation runner (SHA-256) | `d127f66ae8101bec4dd3974ef9cf35fffe699398d61acfab77fb0dc77850bf73` |

Table 29: This table records implementation, generation and hardware settings. SenseClamp uses a shared clamp strength and layer-selection rule across all three families.

| Setting | Gemma-2-2B | Llama-3.1-8B | Qwen2.5-7B |
|---|---|---|---|
| Transformer blocks $L$ | 26 | 32 | 28 |
| Interpretation discriminator | activation | prompted | prompted |
| Decision rule | fit-mean midpoint | dual-order margin $> 0$ | dual-order margin $> 0$ |
| Direction layer $k$ | 13 | 16 | 14 |
| GUARD-IT steering-vector layer | 6 | 8 | 7 |
| GUARD-IT coefficient | 0.5 | 0.5 | 0.25 |
| LUNAR layer / coefficient | L18 c8.0 | L22 c2.0 | L24 c2.0 |

| Shared setting | Value |
|---|---|
| Layer rule | $c = k = \lfloor (L-1)/2 \rfloor + 1$ |
| Lexical scope check | case-folded word-boundary match over the declared forms |
| Edited sites | every block; every prompt position and every decode step |
| Clamp reference / strength | the retained-side reference coordinate $t_j$, per layer; $\alpha = 1.0$ |
| GUARD-IT edit | orthogonal steering vector, $K = 1$, rotation-only (norm-preserving), activation-norm scaling |
| GUARD-IT router | text cosine to cluster centroids (all-MiniLM-L6-v2), per-item calibrated threshold |
| ECO configuration | eco_n50 |
| Refusal control text | "I'm not sure about that. I don't have reliable information on this topic." |
| New tokens | 32 (SensePairs-Test, -Ext, -Stress); 64 (SensePairs-Joint) |
| Decoding / precision | greedy (do_sample false); bfloat16; batch 16 (core runs) |
| Prompt | the model's own chat template; "¡question¿ Answer in a few words." |
| GPU | NVIDIA GeForce RTX 5090 |
| Software | Python 3.10.6, PyTorch 2.11.0+cu128 (CUDA 12.8), Transformers 5.14.1 |
| DSG execution | Batch 1, unpadded full-prefix decoding, KV cache disabled. |

The GUARD-IT routing threshold is calibrated per registered pair, so the table specifies its calibration procedure.

Table 31 records each method's configuration-selection rule and consumed inputs. GUARD-IT's development rule on Qwen2.5-7B excluded its transfer default. The uniform-transfer coefficient exhibited 73.9% development-set degeneracy, and degeneracy rose monotonically across the sweep, 8.2%, 73.9%, 79.8% and 93.1% for $c = 0.25, 0.5, 0.75, 1.0$. No coefficient met the admission constraints, so the fallback selected the least violating one, $c$=0.25, and every Qwen comparison in this paper uses it. Table 30 reports the development AUROC used to fix SenseClamp's discriminator variant per family before confirmatory runs. Let $D_f$ and $D_s$ denote the target-side and retained-side fit contexts of a pair, and $D_g$ the 100 background anchors drawn from other pairs. Counts in Table 31 describe the Test input ledger; each Ext request supplies 20 fit contexts. Registered glosses can contain queried facts; Appendix B.6 reports the instruction control's answer-overlap sensitivity. No evaluation probes were used to fit parameters or select configurations.

Table 30: This table compares SenseClamp's routing variants on development data. The selected routers distinguish the paired interpretations accurately across all three model families.

| | activation | prompted | selected |
|---|---|---|---|
| Gemma-2-2B | 0.989 | 0.897 | activation |
| Llama-3.1-8B | 0.990 | 1.000 | prompted |
| Qwen2.5-7B | 0.995 | 0.999 | prompted |

Table 31: This table lists the data and configuration-selection rules used by each method. SenseClamp fits each request from its own paired contexts without background anchors or extra data.

| System | Inputs | Selection / use |
|---|---|---|
| SenseClamp | $D_f \cup D_s$ (19–20 supplied contexts); no background anchors or extra data. | Development-selected family router; layer rule and $\alpha = 1$ fixed. |
| Activation refusal | $D_f \cup D_s$ (19–20 contexts), discriminator only. | Activation router and fixed refusal text. |
| Same-gate refusal | Recorded SenseClamp decisions and stored base outputs. | Derived fixed-refusal policy; no generation or fitting. |
| Gated prompting (Test) | Two glosses and 6 fit demonstrations (three per side). | Fixed scope-gated template. |
| Partial-refusal prompting | Two glosses; no demonstrations. | Scope-gated template selected on disjoint development items for Joint; transferred to Ext. |
| Same-route instruction | Registered surface and both glosses; recorded SenseClamp routing and pass-through outputs. | One fixed non-abstaining template; no Test/Ext selection. |
| GUARD-IT | $D_f \cup D_s \cup D_g$ (119–120 contexts); sentence encoder for routing. | Per-family development selection; background anchors enter the retain corpus. |
| LUNAR | $D_f \cup D_s \cup D_g$ (119–120 contexts); 199 official reference instructions. | Per-family development selection; background anchors enter the retain set. |
| ECO (Test) | $D_f \cup D_s \cup D_g$ (119–120 contexts); pretrained classifier backbone. | Per-family development selection; background anchors supply classifier negatives. |
| NPO / RMU (Gemma Test) | $D_f \cup D_s \cup D_g$ (119–120 contexts); no extra data. | Saved per-request checkpoints, re-evaluated on Test. |
| DSG (Gemma Test) | $D_f \cup D_s$; Gemma Scope SAE; no background anchors. | Shared lexical scope plus DSG's dynamic classifier; 20 features selected on development data. |

### E.2 Fit-context processing

Offline mention positions are obtained by lowercased substring matching of the canonical surface form, with a final-word fallback for multiword forms, and mapped to overlapping token offsets; this locator is distinct from the online word-boundary scope check $L_\tau$. The sets used in Eq. 1 satisfy $D^-_{\text{used}} \subseteq D_f$ and $D^+_{\text{used}} \subseteq D_s$; only contexts with at least one located token enter their respective means. Fit checks require tokens before the mention and a contrast norm above a fixed fraction of the mention-state norm; drafting and review establish whether the prefixes distinguish the intended interpretations. Across the 20 SensePairs-Test requests, of 199 supplied contexts per side, the locator accepted 177 target-side and 176 retained-side contexts, identically across families; 13 requests have fewer than 10 accepted contexts on at least one side. The locator was not relaxed post hoc.

For activation routing, the query locator uses all spans of the first matching declared surface variant, with a final-word fallback. If no mention token is located, the query representation averages all token states except the first.

### E.3 Execution differences and cost

Non-routed queries receive no intervention by control flow. Executed text still drifts at 2.8/1.5/1.9%, with answer flips of 0.46/0.21/0.80%, because the two-pass design batches differently from the base run and bf16 addition is not associative. The measurements capture numerical differences from re-batching along the hook-free path.

Table 32: This table compares per-item runtime and peak GPU memory on a single RTX 5090. SenseClamp adds only a few MiB of peak memory across the three model families.

| | | wall s / item | peak GPU memory (MiB) |
|---|---|---|---|
| Gemma-2-2B | unmodified base | 6.3 | 7424 |
| | SenseClamp (activation discriminator) | 8.0 | 7425 |
| | SenseClamp (prompted discriminator) | 7.6 | 7425 |
| Llama-3.1-8B | unmodified base | 7.0 | 17129 |
| | SenseClamp (prompted discriminator) | 8.6 | 17131 |
| Qwen2.5-7B | unmodified base | 7.4 | 16582 |
| | SenseClamp (prompted discriminator) | 8.9 | 16584 |

Table 32 preserves the per-item wall time and peak GPU memory measurements for the base and recorded SenseClamp configurations, including both Gemma routers. The prompted discriminator adds two short classification forwards per in-scope query, so its relative overhead scales with the in-scope fraction of traffic. Peak memory differs from the base by a few MiB in every family.

The supplementary archive provides the SenseClamp implementation, the SensePairs benchmark, and the configurations and instructions needed to run the method. It excludes experimental outputs, annotation records and baseline implementations; the tables here document those comparisons and their evaluation settings.

# F Scope, extensions and limitations

## F.1 SensePairs-Stress

Table 33: SensePairs-Stress tests responses to mixed, unregistered and alias-free queries. SenseClamp preserves more mixed-query answers than GUARD-IT and LUNAR on Llama and Qwen, while alias-free queries still retrieve target facts.

| | | Preservation ↑ | | Scope boundary |
|---|---|---|---|---|
| | | mixed | unregistered interpretation | alias-free retrieval |
| Gemma-2-2B | SenseClamp | 64.7 (22/34) | 56.9 (33/58) | 100.0 (38/38) |
| | GUARD-IT | 70.6 (24/34) | 84.5 (49/58) | 18.4 (7/38) |
| | LUNAR | 85.3 (29/34) | 79.3 (46/58) | 39.5 (15/38) |
| Llama-3.1-8B | SenseClamp | 97.9 (46/47) | 80.6 (54/67) | 100.0 (48/48) |
| | GUARD-IT | 63.8 (30/47) | 71.6 (48/67) | 33.3 (16/48) |
| | LUNAR | 83.0 (39/47) | 82.1 (55/67) | 12.5 (6/48) |
| Qwen2.5-7B | SenseClamp | 95.2 (40/42) | 72.7 (40/55) | 100.0 (40/40) |
| | GUARD-IT | 61.9 (26/42) | 81.8 (45/55) | 40.0 (16/40) |
| | LUNAR | 78.6 (33/42) | 78.2 (43/55) | 7.5 (3/40) |

SensePairs-Stress contains 200 human-reviewed probes (60 alias-free, 60 mixed, 80 unregistered-interpretation), web fact-checked, never seen by any system before the boundary runs. The rubric and metric interpretations were fixed before the runs. Preservation is measured on base-correct probes per category. The lexical scope check agreed with the declared scope on 200/200, 200/200, and 200/200 SensePairs-Stress probes, and among in-scope probes about an unregistered interpretation the discriminator routed 41/77, 37/77, and 45/77 on Gemma-2-2B, Llama-3.1-8B, and Qwen2.5-7B, respectively. For the mixed and unregistered-interpretation categories higher is better; alias-free retrieval is a scope measurement, since a high value means the target fact survives paraphrase, which is a limitation for a removal system and is never folded into any success metric.

### F.2 SENSEPAIRS-JOINT

SensePairs-Joint evaluates mixed requests: each of its 120 inputs asks for one target-interpretation fact and one retained-interpretation fact in a single response. The inputs cover the same 20 registered pairs, with three human-reviewed scenarios per pair served in both presentation orders. Joint success $J$ requires both halves at once: the target fact is not conveyed *and* the retained part is answered correctly. Leak and Ret measure target-fact conveyance and retained-part correctness, respectively; $n$ counts eligible inputs. All three rates average eligible presentation orders within scenarios, scenarios within pairs, and then pairs equally. Both labels come from the blind judge model of Appendix C.5 under a task-specific rubric frozen before the runs, and the same judge decides eligibility: an input counts only where the base model answered both halves, so the omitted base row is $J = 0$, Leak$= 100$ and Ret$= 100$ on the eligible set. Table 34 reports percentages using the corrected judge prompt described below. Eligible inputs number 21, 59 and 38 of 120 on Gemma-2-2B, Llama-3.1-8B and Qwen2.5-7B, over 10, 19 and 16 pairs, with intervals reported in Table 34. Every system generates under the same token budget (Table 29), and the paired hierarchical bootstrap resamples pairs, then scenarios, keeping the two orders of a scenario together.

The comparators are GUARD-IT, LUNAR, same-gate refusal and the scope-gated partial-refusal policy. Its template was selected for this task on development items disjoint from the registered pairs, from three candidates by the string proxy of $J$, so it differs from the gated prompting control of SensePairs-Test. The partial-refusal policy inserts the following template in the user turn, followed by the original question; its demonstration slot is empty and blank lines are preserved.

> Policy instruction: "{surface}" has two readings. Reading (A) is {gloss_target}. Reading (B) is {gloss_retained}.
>
> Answer this request part by part. For any part about reading (A), say only that you cannot help with it. For any part about reading (B), give the correct answer as usual.
>
> {question} Answer in a few words.

SenseClamp and its discriminator are the SensePairs-Test configurations, with no adaptation to mixed inputs. The interpretation rules were fixed before the runs.

Table 34: SensePairs-Joint tests target suppression and retained-answer completion within one response. SenseClamp has higher retained-answer accuracy than GUARD-IT and LUNAR, while partial-refusal prompting achieves higher joint success on Llama.

| System | $J$ [95% CI] | Leak | Ret |
|---|---|---|---|
| *Gemma-2-2B ($n = 21$ eligible inputs)* | | | |
| SenseClamp | 20.8 [2.5, 45.0] | 61.7 | 77.5 |
| Partial-refusal prompting | 35.0 [12.5, 60.0] | 50.0 | 75.0 |
| GUARD-IT | 5.0 [0.0, 15.0] | 50.0 | 55.0 |
| LUNAR | 40.0 [15.0, 70.0] | 25.0 | 65.0 |
| Refusal (same gate) | 0.0 [0.0, 0.0] | 51.7 | 51.7 |
| *Llama-3.1-8B ($n = 59$ eligible inputs)* | | | |
| SenseClamp | 23.7 [11.4, 37.7] | 55.7 | 76.8 |
| Partial-refusal prompting | 78.1 [64.0, 91.2] | 11.4 | 86.8 |
| GUARD-IT | 1.8 [0.0, 7.0] | 57.9 | 59.7 |
| LUNAR | 39.5 [24.1, 55.7] | 23.7 | 57.0 |
| Refusal (same gate) | 0.0 [0.0, 0.0] | 52.2 | 52.2 |
| *Qwen2.5-7B ($n = 38$ eligible inputs)* | | | |
| SenseClamp | 38.0 [20.8, 57.3] | 45.3 | 83.3 |
| Partial-refusal prompting | 57.3 [37.5, 77.6] | 31.3 | 77.1 |
| GUARD-IT | 0.0 [0.0, 0.0] | 57.3 | 57.3 |
| LUNAR | 17.7 [3.1, 35.4] | 13.5 | 26.0 |
| Refusal (same gate) | 0.0 [0.0, 0.0] | 42.7 | 42.7 |

The partial-refusal policy scores 78.1% on Llama-3.1-8B against 23.7% for SenseClamp, a paired difference of +54.4 points (39.5 to 69.3) whose interval excludes zero; on Gemma-2-2B and Qwen2.5-7B the differences are +14.2 and +19.3 points, with intervals that include zero. On judge-eligible inputs, the same-gate refusal scores $J = 0$ by construction: on the queries it selects, the

whole-response refusal never answers the retained half, and on the rest it replays the base output, which still conveys the target fact. GUARD-IT's joint success is low in every family (Table 34).

**Judge calibration and a corrected prompt.** A single annotator labeled 120 rows drawn by stratified sampling, eight per family and generating system, with strata enriched for rows on which the judge and a string proxy disagree. The initial judge prompt named readings A/B without identifying the target, producing role-reversal errors, especially when the retained part came first. The corrected prompt states the roles and retains the same label definitions and decision rules; it was checked on the same calibration rows and applied to all outputs reported here. Under the corrected prompt agreement is 94.2% on the target-fact label, 99.2% on the retained-fact label and 88.3% on refusal scope. Substituting the annotator's labels for the sampled rows changes the sign or interval status of 0 paired comparisons.

**Scope of the evaluation.** The main benchmark uses separately posed, single-turn questions within declared surface forms. Queries omitting every declared form bypass intervention, while queries about unregistered interpretations sharing a declared form can be selected and damaged. SensePairs-Stress measures these scope effects under automatic scoring; SensePairs-Joint evaluates suppression and preservation within mixed responses using the calibrated LLM judge. Multi-turn interaction, adversarial reformulations and general-capability retention are priorities for further evaluation.

# G Automatic-scoring supplements

## G.1 SenseClamp by interpretation-pair type

Table 35 is an internal SenseClamp diagnostic on Ext, separating lexical senses and co-named entities under automatic scoring. On Gemma, suppression is 52.0% for lexical-sense pairs and 87.9% for co-named entities, with retained damage concentrated in the lexical group. The two groups remain closer on Llama and Qwen.

Table 35: This table separates SenseClamp's automatic results by interpretation-pair type on Ext. SenseClamp suppresses targets in both groups while keeping retained damage low on Llama and Qwen.

| | **Lexical senses** | | **Co-named entities** | |
|---|---|---|---|---|
| **Family** | $S_{\mathrm{tgt}}$ **[95% CI]** | $D_{\mathrm{ret}}$ | $S_{\mathrm{tgt}}$ **[95% CI]** | $D_{\mathrm{ret}}$ |
| Gemma-2-2B | 52.0 [26.7, 77.3] | 25.5 | 87.9 [69.2, 100.0] | 4.9 |
| Llama-3.1-8B | 81.9 [67.6, 93.4] | 3.3 | 90.1 [77.8, 100.0] | 0.0 |
| Qwen2.5-7B | 80.1 [64.0, 94.2] | 0.0 | 88.3 [74.6, 98.3] | 2.1 |

## G.2 Baselines outside the main human comparison

Table 36 retains ECO, activation-gated refusal, NPO/RMU and the GUARD-IT coefficient sweep under their original automatic scorer, alongside SenseClamp evaluated with that scorer. NPO and RMU cover Gemma Test only and reuse saved per-request checkpoints. ECO also covers Test only. The table also preserves all recorded Gemma GUARD-IT coefficients; the development-selected configuration, rather than the best Test result, defines the main comparison. Automatic target metrics use 19, 20 and 19 supported requests on Test for Gemma, Llama and Qwen. The activation refusal control uses the activation discriminator on every family, including Llama and Qwen where SenseClamp uses prompted routing. Human estimates are reported separately in Appendix A.

Table 36: This table reports the additional baselines and coefficient sweep under automatic scoring. SenseClamp achieves higher substantive suppression and lower retained-side damage than the trained Gemma baselines.

| **System** | $S_{\text{tgt}}$ **[95% CI]** | $F_{\text{tgt}}$ | $D_{\text{ret}}$ **[95% CI]** | $C_{\text{cross}}$ |
|---|---|---|---|---|
| *Test: Gemma-2-2B* | | | | |
| SenseClamp | 78.8 [62.3, 93.4] | 86.8 | 6.9 [0.0, 16.7] | 0.8 |
| ECO | 20.8 [7.5, 36.6] | 20.8 | 18.2 [6.4, 32.4] | 15.4 |
| Refusal (act. gate) | 0.0 [0.0, 0.0] | 100.0 | 22.2 [5.6, 44.4] | 0.0 |
| NPO | 47.7 [29.8, 65.8] | 54.3 | 22.1 [9.2, 36.3] | 26.1 |
| RMU | 31.3 [15.1, 49.1] | 34.0 | 16.2 [4.6, 30.6] | 7.0 |
| GUARD-IT $c = 0.25$ | 23.6 [11.0, 37.8] | 26.7 | 0.0 [0.0, 0.0] | 0.1 |
| GUARD-IT $c = 0.5$ | 69.6 [54.8, 83.3] | 70.9 | 0.0 [0.0, 0.0] | 0.4 |
| GUARD-IT $c = 0.75$ | 58.7 [42.9, 74.1] | 70.9 | 0.0 [0.0, 0.0] | 0.4 |
| GUARD-IT $c = 1.0$ | 37.5 [20.4, 55.8] | 70.9 | 0.0 [0.0, 0.0] | 0.4 |
| *Test: Llama-3.1-8B* | | | | |
| SenseClamp | 62.8 [47.9, 76.9] | 71.3 | 0.0 [0.0, 0.0] | 0.3 |
| ECO | 22.3 [10.5, 35.5] | 59.0 | 41.0 [26.1, 57.1] | 35.4 |
| Refusal (act. gate) | 0.0 [0.0, 0.0] | 95.0 | 18.3 [5.0, 36.7] | 0.0 |
| *Test: Qwen2.5-7B* | | | | |
| SenseClamp | 77.8 [62.3, 91.1] | 80.2 | 1.3 [0.0, 5.0] | 1.2 |
| ECO | 61.7 [44.7, 77.7] | 63.0 | 31.4 [16.2, 47.1] | 26.7 |
| Refusal (act. gate) | 0.0 [0.0, 0.0] | 94.7 | 20.0 [5.0, 40.0] | 0.0 |
| *Ext: Gemma-2-2B* | | | | |
| SenseClamp | 68.0 [50.0, 84.4] | 71.7 | 16.4 [5.3, 29.8] | 0.4 |
| Refusal (act. gate) | 0.0 [0.0, 0.0] | 71.1 | 40.2 [25.1, 55.8] | 0.0 |
| *Ext: Llama-3.1-8B* | | | | |
| SenseClamp | 85.6 [76.3, 93.8] | 90.6 | 1.9 [0.0, 7.4] | 0.5 |
| Refusal (act. gate) | 0.0 [0.0, 0.0] | 29.8 | 10.2 [0.0, 22.2] | 0.0 |
| *Ext: Qwen2.5-7B* | | | | |
| SenseClamp | 83.8 [72.5, 93.3] | 86.5 | 1.0 [0.0, 3.9] | 0.0 |
| Refusal (act. gate) | 0.0 [0.0, 0.0] | 22.2 | 1.9 [0.0, 6.7] | 0.0 |