A Pre-Generation Epistemic Routing Axis in Language Models

Steve Bakos[1] **Heidar Davoudi**[2] **Aijun An**[1]
[1]Department of Electrical Engineering and Computer Science, York University, Toronto, Canada
[2]Faculty of Science, Ontario Tech University, Oshawa, Canada
{sbakos,aan}@yorku.ca
heidar.davoudi@ontariotechu.ca

Abstract

When answering factual questions, language models may respond correctly, hallucinate, or abstain. We investigate whether this epistemic disposition is encoded in hidden states before any answer token is generated. Using a balanced Wikipedia QA corpus split by article date, we retain questions for which eight independently sampled answers receive the same label. We then analyze prompt-only activations from the multilayer perceptron (MLP) blocks of four instruction-tuned 7–9B models. Principal Component Analysis (PCA) reveals a dominant pre-generation epistemic routing axis ordering ABSTAIN, HALLUCINATION, and CORRECT. As a bounded causal test, we edit along the axis-aligned ABSTAIN–HALLUCINATION contrast and compare answers to each question before and after. Editing the MLP outputs across a fixed layer range shifts many hallucinations chiefly to abstention, but effects vary by model; restricting the edit to selected MLP neurons is less reliable. The paired audit distinguishes useful changes from correct loss and broken outputs, motivating conditional controllers.

1 Introduction

Large language models are now routinely used for question answering, but they still face a basic epistemic failure: when uncertain or missing a fact, they may produce a *fluent hallucination* instead of saying they do not know. A useful model should preserve known answers, avoid hallucinated answers, and *abstain* when the answer is not reliable.

Existing work shows that hidden states encode truthfulness, uncertainty, and hallucination risk (Chen et al., 2024; Duan et al., 2024; Orgad et al., 2025; Gao et al., 2025); representation engineering further shows that some behaviors, including hallucination, can be shifted by *activation steering*, an inference-time activation edit that leaves model weights unchanged and modifies internal hidden states (Zou et al., 2023; Turner et al., 2024; Marks and Tegmark, 2024). Recent QA and abstention benchmarks increasingly distinguish correct answers, hallucinations, and abstentions (Kirichenko et al., 2025; Feng et al., 2024; Mohamadi et al., 2025); our question is whether those outcomes also form stable, model-specific activation classes before generation. The main challenge is that the model outcomes are entangled. That is, a model may internally separate uncertainty from answering behavior while still placing correct and hallucinated answers close together. In that case, a steering vector that reduces hallucinations by pushing toward abstention can also erase correct answers.

The central question is therefore not only whether a model contains a diagnostic signal, but whether that signal can support behaviorally useful interventions: reducing hallucinations without sacrificing answers that were already correct. We study this question at the pre-generation decision point where the model has read the question but before it generates any answer token. This avoids a common ambiguity in hallucination studies, where post-generation signals may reflect response length, wording, or decoding artifacts.

We organize the empirical analysis around four questions: whether pre-generation activations organize CORRECT, HALLUCINATION, and ABSTAIN outcomes, and whether this structure recurs across model families and activation spaces; whether an ABSTAIN-minus-HALLUCINATION (A-H) activation edit reduces hallucinations without excessive collateral damage; when fixed activation steering becomes brittle; and whether the signal reflects a pre-answer knowledge state rather than surface-level text artifacts.

To enable this investigation, we call each evaluated model the *solver* and the separate model that labels its answers the *judge*. For each solver, *8/8 consensus* means that all eight sampled answers receive the same judged label; its *activa-*

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 28066–28091
October 24-29, 2026 ©2026 Association for Computational Linguistics

tion bank retains the corresponding pre-answer hidden states. We build banks for Qwen-2.5-7B, Mistral-7B, Llama-3.1-8B, and Gemma-2-9B using Wikipedia QA questions answered under a fixed prompt. We capture each hidden state at the last prompt token, before generation, and use it for analysis and steering.

PCA on these activation banks reveals a dominant pre-generation axis that orders the ABSTAIN, HALLUCINATION, and CORRECT centroids. We refer to this direction as an *epistemic routing axis*, as it organizes the model's pre-generation disposition to abstain, hallucinate, or answer correctly. We show that this structured geometry is not an isolated idiosyncrasy, but it consistently recurs in both the multilayer perceptron (MLP) output written back into the residual stream (*MLP residual-write*) and the post-gating MLP state before that write (*MLP-intermediate*) of these four open-weight model families.

The observed geometry shows that pre-generation hidden states contain a diagnostic signal aligned with the A–H steering direction, while also making the risk of collateral damage explicit. In several models, HALLUCINATION and CORRECT remain close relative to the large ABSTAIN displacement. This defines the central intervention risk: the same fixed edit that moves hallucinations toward abstention can also move baseline-correct answers away from correctness.

Contributions. This paper makes three contributions. **(1)** We identify a pre-generation epistemic routing axis across four model families and two activation spaces. **(2)** We derive fixed edits from the axis-aligned A-H contrast, then use paired transition scoring to test when they are useful, lossy, or brittle. **(3)** We introduce a reusable pre/post-cutoff QA construction procedure: choose a cutoff date, build likely-known and likely-missing question pools around it, label each solver model's own rollouts, and keep separate steering, validation, and test datasets. For later models, the same procedure can be reused by changing the cutoff date.[1]

2 Related Work

Hallucination, abstention, and hidden-state signals. Hallucination evaluation spans imitative falsehoods, freshness, long-tail memory, and open-domain QA (Lin et al., 2022; Vu et al., 2024; Mallen et al., 2023; Kwiatkowski et al., 2019). Work on semantic uncertainty, self-knowledge, and multi-sample consistency shows that unstable generations expose hallucination risk (Kuhn et al., 2023; Farquhar et al., 2024; Manakul et al., 2023; Kadavath et al., 2022; Amayuelas et al., 2024); Calibration theory shows why hallucinations can remain unavoidable (Kalai and Vempala, 2024); abstention work instead studies withholding answers when knowledge is missing (Kirichenko et al., 2025; Feng et al., 2024; Mohamadi et al., 2025). Our setting therefore needs three epistemic outcomes, not a binary true/false probe: supported answer, unsupported answer, and abstention.

Several papers show that internal activations encode truthfulness, answerability, or hallucination signals before or during generation (Chen et al., 2024; Gottesman and Geva, 2024; Slobodkin et al., 2023; Duan et al., 2024; Gao et al., 2025). Recent work also identifies entity-level knowledge-awareness features and question-only signals of answer accuracy (Ferrando et al., 2025; Moreno Cencerrado et al., 2026). Local intrinsic dimension (LID) distinguishes truthful from untruthful completed generations (Yin et al., 2024); by contrast, we study final-prompt-token states before generation and distinguish three later outcomes: correct answers, hallucinations, and abstentions. A prompt-token, three-outcome LID adaptation remains open. Work interpreting prompt injection as role confusion further shows that role-tag formatting can influence how a model interprets instructions (Ye et al., 2026), motivating our prompt-format control. Work distinguishing ignorance from error further motivates the use of likely-known pre-cutoff and likely-missing post-cutoff source pools (Simhi et al., 2024; Lu et al., 2026). These results motivate looking inside the model, but they do not by themselves show whether a pre-generation signal can be edited without collateral damage.

Activation steering and causal audits. Representation engineering and activation-addition methods show that simple activation directions can alter model behavior without weight updates (Zou et al., 2023; Turner et al., 2024; Rimsky et al., 2024; Li et al., 2023; Marks and Tegmark, 2024). Adaptive and conditional steering work emphasizes that intervention reliability depends on when, where, and how strongly a direction is applied (Wang et al., 2025; Lee et al., 2025; Da Silva

[1] Code, data, prompts, and reproducibility artifacts are available in the public artifact repository.

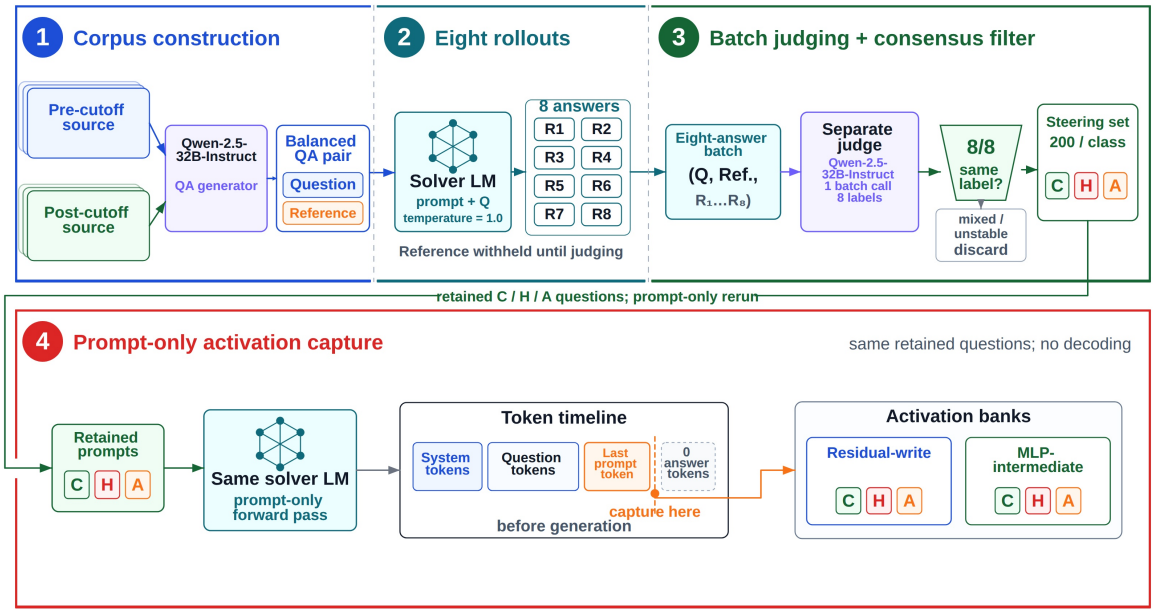


Figure 1: **Corpus-to-activation pipeline.** Pre/post-cutoff QA pairs yield eight solver answers without the reference. One judge receives all eight together and labels each answer. We retain 8/8-consensus questions, sampling 200/class. Prompt-only reruns capture final-token residual-write and MLP-intermediate activations.

et al., 2025; Bas and Novak, 2025). Refusal steering is related because it controls refusal behavior (Arditi et al., 2024), but our target is epistemic QA abstention rather than safety refusal. Dense Window applies layer-specific A–H edits across a validation-selected consecutive residual-write range. The sparse comparator applies the raw MLP-intermediate A–H edit on a fixed mask. CETT (cumulative errors of tail truncation), adapted by H-Neurons, scores neurons to select that mask but does not scale the edit (Zhang et al., 2024; Gao et al., 2025). Section 4.2 defines both interventions. The question is when the axis is editable, lossy, or too brittle for fixed steering.

Positioning. Closest prior work supplies the components: hidden-state probes show diagnostic signals, pre-generation studies show knowledge/answerability information before decoding, and steering work shows that directions can change behavior. The missing piece is the combination we test here: a three-outcome pre-generation geometry over Correct, Hallucination, and Abstain, paired with a transition audit that rewards repair while exposing lost correct answers and unusable outputs.

3 Constructing Activation Banks

Our first methodological step is to turn factual QA behavior of solver models into model-specific activation banks. The QA task has three response types: given a question, a solver model may produce a supported answer (CORRECT), an unsupported factual answer (HALLUCINATION), or an epistemic non-answer (ABSTAIN). These labels are not imported from an off-the-shelf benchmark. They are produced from each solver model's own rollouts under the same primary XML-formatted prompt (Appendix A), because the activation classes must describe the behavior of the model whose hidden states are being analyzed. An activation bank is the retained collection of the solver's pre-answer hidden states, grouped by the judge-assigned outcome.

Figure 1 shows the four-stage procedure. Stage (1) builds the QA pool used for the later solver rollouts. We collect English Wikipedia source passages and record each source article's creation date. We split passages into a pre-cutoff pool, from articles created before January 1, 2024, and a post-cutoff pool, from articles created after January 1, 2024. We then ask a Qwen-2.5-32B-Instruct generator to convert each source passage into one factual question plus a short reference answer supported by that passage. After QA generation and parsing, we balance the two pools: the post-cutoff pool is smaller after filtering, so all 3,426 usable post-cutoff QA pairs are kept and 3,426 pre-cutoff QA

pairs are sampled to match.[2]

The January 1, 2024 split is an operational corpus construction heuristic, not a claim that all solver models share the same parametric knowledge cutoff and not a label assignment. The date immediately follows Llama-3.1's reported December 2023 cutoff and gives a single auditable boundary for constructing likely-known and likely-missing source pools across model families whose exact training exposure is not public or directly comparable (Meta, 2024; Yang et al., 2024; Jiang et al., 2023; Gemma Team, 2024; Cheng et al., 2024). Accordingly, pre-cutoff questions are more likely to concern facts the model knows and post-cutoff questions facts it does not, but either side can still be judged as CORRECT, HALLUCINATION, or ABSTAIN (Simhi et al., 2024). This new QA pool is needed because existing ternary benchmarks do not simultaneously provide source passages, reference answers, article-creation cutoff information, and labels induced from each solver model's own rollouts.

In stage (2), each solver model produces 8 independently sampled responses to every balanced-corpus question at temperature 1.0; we call each response a rollout. The solver sees only the prompt template and question. The Qwen-2.5-32B-Instruct judge receives the eight rollouts in one batch, together with the question and reference answer, and assigns one label to each; it is not told whether the question came from the pre- or post-cutoff pool. The pre/post field is retained for balancing and split audits, not used by the judge. In stage (3), we keep a question only if all 8 rollouts receive the same one of the three labels: CORRECT, HALLUCINATION, or ABSTAIN. Mixed or unstable questions are discarded to favor label certainty over coverage. We refer to this as the 8/8-consensus criterion.

The equal-sized pre/post pools produce different supplies under this criterion: pre-cutoff questions supply stable CORRECT controls, post-cutoff questions supply stable ABSTAIN examples, and stable HALLUCINATION examples occur in both. Table 1 reports the model-specific counts.

In stage (4), retained 8/8-consensus questions are rerun in a prompt-only forward pass. We snapshot *MLP residual-write* activations (the down-projection output written back to the residual stream) and *MLP-intermediate* (post-gating, pre-down-projection) activations at the last prompt token before the model generates its first answer token (Gottesman and Geva, 2024; Duan et al., 2024). For each question and activation space, we concatenate this last-token state across MLP layers into one all-layer activation vector. Each model's *steering dataset* has 200 8/8-consensus examples per class (C/H/A). PCA uses all 600 activations; A and H means define layer-specific A–H vectors. Validation selects hyperparameters; held-out and external data are evaluation-only. Neither enters PCA nor vector construction; see Section 5 and Appendix B.

[2]Filtering criteria are listed in Appendix B.

Table 1: **Why both cutoff pools are needed.** Each era contains 3,426 candidate questions. Counts are questions for which all eight solver rollouts receive the same label. Pre-cutoff supplies more stable CORRECT controls, post-cutoff supplies more stable ABSTAIN examples, and HALLUCINATION occurs in both. Era enriches the pool but does not assign its labels.

Model	Era	C	H	A
Qwen-2.5-7B	pre	680	1066	52
Qwen-2.5-7B	post	77	1126	280
Mistral-7B	pre	507	602	147
Mistral-7B	post	55	386	780
Llama-3.1-8B	pre	578	707	42
Llama-3.1-8B	post	67	588	356
Gemma-2-9B	pre	754	608	412
Gemma-2-9B	post	108	542	1073

4 Analyzing and Intervening on Activation Banks

Once the model-specific activation banks are constructed, the remaining method has three steps: first, identify the activation geometry in the banks; second, turn that geometry into intervention candidates; and third, audit whether those candidates repair hallucinations without damaging correct answers.

4.1 Activation Geometry

For each model, we use Principal Component Analysis (PCA) (Jolliffe and Cadima, 2016) on *MLP residual-write* (down-projection) and *MLP-intermediate* activations from the last prompt position, and then use the activation labels to interpret the projected centroids. Figure 2 shows the residual-write projections onto the first two principal components (PC1 and PC2); Appendix Figure 6 gives the corresponding MLP-intermediate projections. For each setting, we report the percentage of variance explained by PC1 and mark the class centroids. Across all eight settings, the same cen-

Table 2: **Compact pre-generation geometry.** RW and Int. denote residual-write and MLP-intermediate spaces. $\mathrm{A-H/PC1}\,|\cos|$ compares the supervised A–H contrast with PC1; near one means aligned. Distance ratios use the full activation vectors; brackets give percentile 95% CIs, all below one, placing hallucination closer to correct than to abstention.

Model	Space	PC1 var. (%)	A–H/PC1 $\lvert\cos\rvert$	$d(H,C)/d(H,A)$ [95% CI]
Qwen	RW	40.6	0.98	0.479 [0.410, 0.574]
	Int.	37.9	0.97	0.501 [0.429, 0.595]
Mistral	RW	19.9	0.99	0.273 [0.256, 0.346]
	Int.	19.4	0.99	0.278 [0.260, 0.350]
Llama	RW	63.6	0.99	0.164 [0.140, 0.194]
	Int.	63.4	0.99	0.162 [0.139, 0.192]
Gemma	RW	22.4	0.93	0.225 [0.212, 0.289]
	Int.	22.4	0.94	0.228 [0.213, 0.291]

troid ordering appears on the PC1 dimension: ABSTAIN < HALLUCINATION < CORRECT. In the full activation vectors, $d(H,C)/d(H,A)$ ranges from 0.162 to 0.501, with every 5,000-resample 95% bootstrap interval below one; the absolute distances and intervals are in Appendix Table 14.

One way to reduce hallucination is to steer the activation vectors from HALLUCINATION towards ABSTAIN. This can be done by moving the activations along the direction defined by:

$$\begin{aligned} d_{\mathrm{A-H}} &= \mu_A - \mu_H, \\ v_{\mathrm{A-H}} &= d_{\mathrm{A-H}}/\|d_{\mathrm{A-H}}\|_2. \end{aligned} \tag{1}$$

In Equation 1, μ_A, μ_H are the all-layer class means, $d_{\mathrm{A-H}}$ their raw difference, and $v_{\mathrm{A-H}}$ its unit-normalized form for PC1 comparison; steering uses layer-specific contrasts.

Table 2 reports full-bank $|\cos(v_{\mathrm{A-H}}, \mathrm{PC1})|$; values near one indicate alignment. Appendix Tables 22 and 23 give the corresponding Dense-range and sparse-mask checks.

4.2 Activation Steering Edits

Activation steering modifies the generation process by injecting additional activation vectors without changing the model weights. Correct examples enter the trinary PCA but not the A–H contrast used to define the edit; baseline-correct validation and test answers provide the behavioral guardrail for collateral damage.

Dense Window edits a validation-selected residual-write range; the sparse comparator applies a fixed-mask MLP-intermediate A–H update. CETT scores select that mask but do not scale the edit. Fixed settings apply to every prompt; there is no per-example controller.

Dense Window Steering. Dense Window edits the MLP down-projection output, the residual-write vector a selected layer contributes back to the residual stream. For each layer ℓ in the selected start-to-end layer range L, we compute the A-H difference and normalize it:

$$d_\ell = \mu_\ell^{A,\mathrm{dp}} - \mu_\ell^{H,\mathrm{dp}}, \qquad v_\ell = d_\ell/\|d_\ell\|_2. \tag{2}$$

In Equation 2, $\mu_\ell^{A,\mathrm{dp}}$ and $\mu_\ell^{H,\mathrm{dp}}$ are class-mean down-projection outputs for ABSTAIN and HALLUCINATION examples at layer ℓ. During generation, Dense adds this normalized direction with strength α at each selected layer: $o_\ell^{\mathrm{dp}} \leftarrow o_\ell^{\mathrm{dp}} + \alpha v_\ell$, where o_ℓ^{dp} denotes the current down-projection output. Thus Dense does not broadcast one global vector to all layers. It uses a separate normalized A-H direction for each layer in the selected range, and the same scalar strength α is applied to those selected layers. Appendix J reports the selected layer ranges and their PC1-alignment checks.

Sparse Comparator. The sparse comparator edits the MLP-intermediate space in two steps. First, an L1 probe trained on standardized CETT contribution-score features selects intermediate neurons that distinguish abstention from hallucination on the steering dataset. The nonzero probe coordinates form a fixed mask m_ℓ. Second, steering uses the raw intermediate A-H difference on that mask.

$$u_\ell = \mu_\ell^{A,\mathrm{int}} - \mu_\ell^{H,\mathrm{int}}, \quad z_\ell \leftarrow z_\ell + \alpha\, m_\ell \odot u_\ell. \tag{3}$$

In Equation 3, $\mu_\ell^{A,\mathrm{int}}$ and $\mu_\ell^{H,\mathrm{int}}$ are class-mean MLP-intermediate activations, z_ℓ is the current intermediate activation, m_ℓ is the selected sparse neuron mask, $\odot$ is elementwise multiplication, and α is the edit strength. Unlike Dense, this edit is not normalized separately in each layer; it keeps the raw A-H component sizes on the selected intermediate coordinates and scales the masked edit by α. Appendix I gives the lower-level intervention details, and Appendix J reports the mask-alignment checks.

4.3 Before/After Steering Evaluation

We evaluate steering by comparing the same question before and after the edit. First, the unedited model answers the question and the judge assigns

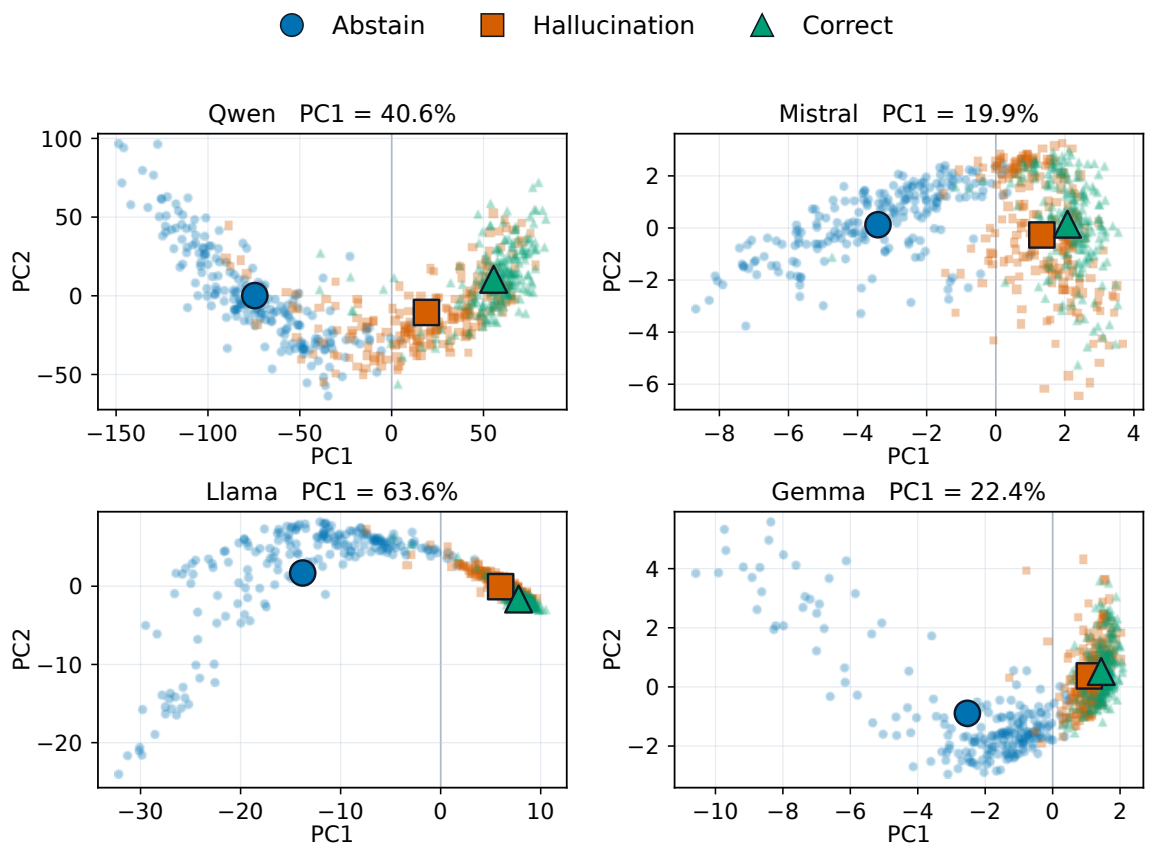


Figure 2: **Pre-generation residual-write PCA projection.** Points show all-layer final-prompt-token residual-write activations; large markers are class centroids. PC1 signs are oriented for display so Abstain–Hallucination–Correct (A–H–C) runs left to right. Appendix Figure 6 shows MLP-intermediate space; Table 2 gives PC1/A–H alignment.

a baseline label. Then the steered model answers the same question and the judge assigns a steered label.

The judge uses Correct (C), Hallucination (H), or Abstain (A); the deterministic Broken-output detector relabels unusable generations as Broken (B) (Appendix G).

Raw paired transitions are primary: baseline H→A is suppression through abstention, H→C is factual recovery, and correct loss plus usable-to-broken transitions measure collateral damage. TransitionScore is a secondary, offline summary that weights source classes equally so the largest class cannot dominate; it compares fixed settings rather than deciding whether to steer an individual question. It assigns credit when an edit moves a hallucination to abstain or correct, or moves an abstention to correct. It assigns a penalty when a baseline-correct answer is lost, when an abstention becomes hallucination or broken, or when a hallucination becomes broken. No-change transitions contribute zero. Figure 3 summarizes this before/after scoring logic. We report raw four-class rates alongside TransitionScore because a single score can hide over-abstention, correct-answer loss, or unusable output (Feng et al., 2024; Mohamadi et al., 2025). Appendix C gives the exact scoring rules.

4.4 Validation Selection

We rank Dense range/strength pairs and fixed-mask sparse strengths on a balanced 128-question validation subset (the *validation-scout*). The top three per family are rerun on the full 500-question validation set; we freeze the highest-scoring setting per model and family for final evaluation. The score supplies damage penalties; no hard threshold is imposed.

5 Empirical Evaluation Setup

Models. We use four instruction-tuned open-weight models as solver models: Qwen-2.5-7B (Yang et al., 2024), Mistral-7B (Jiang et al., 2023), Llama-3.1-8B (Grattafiori et al., 2024), and Gemma-2-9B (Gemma Team, 2024). These models differ in layer count, abstention style, and steering sensitivity, so they test whether the pre-generation

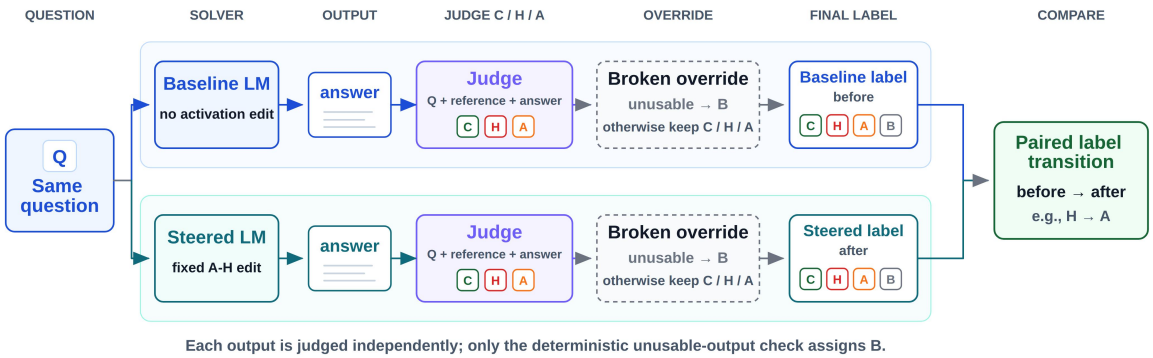


Figure 3: **Steering and before/after audit.** The same solver answers the same question unedited and with the fixed A–H edit. Outputs are judged independently as C/H/A; the BROKEN-output detector assigns B. The final pair defines the transition: H→A and H→C are repairs; correct-answer loss and new B are collateral effects. Appendix C gives the exact TransitionScore weights.

geometry and steering effects recur across model families.

Datasets. The Wikipedia-derived QA suite is organized into named dataset roles. The *steering dataset* is sampled after model-specific 8/8 consensus and contains 200 retained questions per behavioral class (C/H/A) for each solver model; it is used for PCA, activation-bank construction, and steering-vector construction. The *validation dataset* contains 500 questions balanced across pre- and post-cutoff questions (250 each). Its balanced 128-question *validation-scout* subset supports the broad search described in Section 4.4. Final evaluation uses a separate 1,000-question *test dataset* balanced across pre/post cutoff (500 each), plus TruthfulQA, FreshQA, NQ Open, KUQ, and PopQA; Appendix Table 4 gives benchmark sizes and roles.

Steering and Transition Audit. Unlike the batched judging used to establish 8/8 consensus, steering evaluation judges each baseline (unsteered) and steered output separately. Each one-answer prompt contains the question, reference answer, and parsed model answer; Qwen-2.5-32B-Instruct is the judge. For Wikipedia-derived validation/test examples, the stored source passage is also supplied; for external benchmarks, no passage is supplied. The judge first assigns C/H/A. We then apply our deterministic *Broken-output detector*, defined in Appendix G, to the raw response and parsed answer: empty, malformed, repetitive, or degenerate generations are relabeled B. A blind human audit of 100 stratified model outputs agrees with the automated four-class pipeline on 90.0% of examples ($\kappa = 0.864$; Appendix G). As a sensitivity check, three alternate judges relabel the same 12,000 saved answers (four solvers × 1,000 questions × baseline, Dense, and sparse), agreeing with the Qwen judge on 92.3–95.4% of four-class labels. Across 4,000 unsteered answers, class rates differ by at most 3.3 percentage points. Five of 48 Baseline-to-Dense checks (three alternate judges × four solvers × four effects) are too small to call (an absolute TransitionScore of at most 1; an absolute hallucination-rate, abstention-rate, or H→A-minus-H→C repair-gap change of at most 2 percentage points); the other 43 agree with the direction measured by the Qwen judge (Appendix Table 15). Because the questions and solver outputs remain fixed, this isolates judge sensitivity; it does not test Qwen's separate role in generating the Wikipedia-derived QA corpus, which would require rebuilding the corpus with another generator.

6 Analysis

Throughout this analysis, an internal activation state is the pre-generation activation vector captured at the last prompt token. The MLP *residual-write* activations refer to MLP down-projection outputs written back to the residual stream, and *MLP-intermediate* activations are the post-gating, pre-down-projection feed-forward states used by the sparse comparator.

RQ1: Can models encode and organize epistemic disposition in internal states before generating the answer? Figure 2 shows the residual-write projections, and Appendix Figure 6 shows the corresponding MLP-intermediate projections. Across all four models and both activation spaces, PC1 orders the last-prompt-token class centroids as ABSTAIN–HALLUCINATION–CORRECT before

any answer is generated. We call this first principal component the *epistemic routing axis*. Across models, PC1 explains 19.9–63.6% of residual-write variance and 19.4–63.4% of MLP-intermediate variance.

Table 2 quantifies the link to intervention: the global supervised A–H contrast strongly aligns with PC1 ($|\cos(v_{\mathrm{A-H}}, \mathrm{PC1})| = 0.93$–$0.99$). Appendix Tables 22 and 23 audit the same relationship inside the Dense layer range and selected sparse mask. This alignment does not make PC1 the steering direction: PC1 is an unsupervised diagnostic of variation across all three classes, whereas steering uses the supervised A–H contrast to target movement from HALLUCINATION toward ABSTAIN. The geometry therefore suggests where to steer; the paired transition audit tests whether the edit repairs hallucinations without damaging correct answers.

Finding. Before generation, both activation spaces exhibit the same ABSTAIN–HALLUCINATION–CORRECT PC1 order across model families. PC1 diagnoses the geometry; the A–H difference defines the steering edit.

RQ2: Can we manipulate internal states to reduce hallucinations without excessive collateral damage? Equation 1's global all-layer A–H contrast diagnoses PC1 alignment; causal tests instead use the layer-specific Dense and sparse updates in Equations 2 and 3. A useful update should move baseline-HALLUCINATION outputs toward ABSTAIN or CORRECT without excessive CORRECT loss or new BROKEN outputs. Dense Window is primary; the sparse comparator tests whether editing only selected intermediate neurons is sufficient.

After selecting the layer range or mask and edit strength on validation, we freeze each setting for final evaluation. Table 3 reports the primary paired outcomes on every dataset without combining distinct repair and damage pathways. H→A exceeds H→C in 23 of 28 model–dataset evaluations, while correct loss exceeds 20% in 21 of 28. Newly broken outputs are generally rare for Qwen and Gemma, but reach 2.8–8.6% across Llama evaluations and 3.0% for Mistral on KUQ. Thus Dense often changes behavior, but its dominant effect is abstention rather than factual recovery and its collateral cost varies across models and datasets.

As a TransitionScore-independent check, cutoff-proxy F1 treats post-cutoff questions as the positive target and abstention as the positive prediction. It asks whether steering makes abstention better track the coarse pre/post boundary; the cutoff is not an item-level knowledge label. Dense raises this F1 by 11.3–35.0 points for Qwen, Mistral, and Gemma (all paired 95% intervals exclude zero); Llama's $+2.2$-point change $[-1.5, 6.0]$ is inconclusive. The gains come from higher recall while precision falls. An unsteered variant of the primary XML-formatted prompt with an added abstention instruction preserves more correct answers for the former three, whereas Dense suppresses more hallucinations (Appendix Tables 16 and 17).

TransitionScore is retained as a secondary class-balanced summary, not the primary evidence. Across the five external benchmarks, its point estimate is positive for all five Mistral and Gemma evaluations, three of five Qwen evaluations, and none of the five Llama evaluations. Appendix Figure 7 visualizes the pattern; Tables 8–11 give the detailed rates and intervals.

Across the released paired artifacts, the sparse comparator leaves hallucination unchanged as the largest destination in 27 of 28 model–dataset settings. Appendix Table 13 reports its raw transition rates across all datasets.

Finding. Dense can change behavior, but chiefly by increasing abstention and with substantial model-dependent correct loss. Ordered intermediate-space geometry does not by itself guarantee sparse behavioral control.

RQ3: When does fixed activation steering become brittle? Here, brittleness means repair with correct loss or BROKEN outputs. Under the primary XML-formatted prompt, Llama is the clearest brittle case.

Prompt-protocol control. To test whether Llama is unsteerable, we compare a plain direct-answer prompt (D0) with an abstention-aware variant (D1) that instructs "I don't know" when uncertain; both omit XML tags and displayed reasoning. For each, we refit Llama's 200-per-class A–H direction, keep the selected layer window and strength fixed, and test the same held-out 1,000 questions. D0 yields an inconclusive TransitionScore of -0.45 [95% CI -1.70, 1.16], whereas D1 yields a positive TransitionScore of 4.91 [2.00, 8.09], repairing 15.23% of hallucinations with 3.88% correct loss. Appendix F gives the exact prompts; Appendix Table 18 reports all four models.

Table 3: **Dense steering outcomes across datasets.** ID-pre/post are the pre/post-cutoff halves of the held-out in-distribution test set. Each percentage answers a separate question. Of answers that hallucinated before steering, what share became abstentions (H→A) or correct (H→C)? Of answers that were correct, what share stopped being correct (C loss)? Across all questions, what share changed from usable to BROKEN under the Broken-output detector (New B)? Because these starting groups differ, the rows should not be combined. Values are point estimate ± the larger side of a 95% paired-bootstrap interval (5,000 resamples). Exact intervals and counts are provided with the released results.

Model	Outcome	ID-pre	ID-post	Truth	Fresh	NQ	KUQ	Pop
Qwen	H→A	39.9 ± 5.9	80.6 ± 4.4	19.6 ± 4.1	34.9 ± 4.7	20.4 ± 3.2	32.1 ± 3.7	58.0 ± 3.8
	H→C	10.1 ± 3.6	1.3 ± 1.4	25.6 ± 4.3	9.6 ± 3.1	8.5 ± 2.4	3.6 ± 1.6	1.8 ± 1.1
	C loss	37.6 ± 7.2	48.9 ± 15.0	21.7 ± 4.1	27.2 ± 7.1	21.5 ± 4.1	27.4 ± 5.8	34.5 ± 6.5
	New B	0.0 ± 0.0	0.4 ± 0.6	0.2 ± 0.4	0.0 ± 0.0	0.8 ± 0.6	0.4 ± 0.4	0.6 ± 0.5
Mistral	H→A	73.8 ± 5.6	96.5 ± 2.7	45.4 ± 5.3	74.0 ± 4.9	66.0 ± 4.2	66.9 ± 4.1	81.7 ± 3.2
	H→C	8.7 ± 3.6	0.0 ± 0.0	29.1 ± 4.8	5.2 ± 2.6	10.1 ± 2.7	3.5 ± 1.7	2.4 ± 1.4
	C loss	60.2 ± 7.2	78.4 ± 11.8	32.5 ± 4.4	52.7 ± 7.2	51.1 ± 4.5	54.5 ± 6.9	70.9 ± 5.8
	New B	0.0 ± 0.0	0.2 ± 0.4	2.0 ± 1.0	0.2 ± 0.3	0.4 ± 0.5	3.0 ± 1.1	0.7 ± 0.6
Llama	H→A	3.2 ± 2.5	12.7 ± 4.5	0.8 ± 1.0	8.0 ± 3.2	1.0 ± 1.1	7.5 ± 2.2	2.0 ± 1.2
	H→C	8.2 ± 3.8	2.2 ± 2.1	8.7 ± 2.9	4.2 ± 2.5	8.4 ± 2.9	2.6 ± 1.4	4.4 ± 1.8
	C loss	8.3 ± 4.1	9.1 ± 8.4	12.0 ± 3.5	11.5 ± 4.8	9.6 ± 2.7	15.2 ± 5.1	10.0 ± 3.6
	New B	2.8 ± 1.4	8.6 ± 2.4	3.9 ± 1.5	4.3 ± 1.7	3.5 ± 1.2	3.4 ± 1.2	4.9 ± 1.4
Gemma	H→A	54.0 ± 7.0	84.5 ± 5.4	24.9 ± 5.0	64.2 ± 5.8	42.5 ± 4.5	36.0 ± 4.2	64.7 ± 3.9
	H→C	6.2 ± 3.6	0.5 ± 1.1	14.7 ± 4.2	6.6 ± 3.1	3.4 ± 1.7	3.1 ± 1.7	2.2 ± 1.2
	C loss	31.9 ± 6.6	54.4 ± 13.0	21.6 ± 3.9	29.0 ± 5.8	24.8 ± 4.0	27.7 ± 5.8	36.4 ± 5.8
	New B	1.0 ± 1.0	0.8 ± 0.8	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.2 ± 0.3	0.0 ± 0.0

Finding. Fixed steering is model- and protocol-dependent. Llama's D1 result argues against universal failure, but neither control isolates XML formatting as the cause; utility still requires repair to outweigh collateral damage. Appendix Table 6 gives paired examples.

RQ4: Are these findings about knowledge, or are they just surface-level text artifacts? Because the axis is measured before generation, answer wording cannot create it. Pre/post refits preserve the class order in all 16/16 model–space–split fits. All 320/320 class-balanced resamples preserve the order (minimum $n = 14$ per class), showing that the order is not explained by era-driven class imbalance. A blind human audit agrees with the four-class labels on 90.0% of examples ($\kappa = 0.864$). Question-only TF–IDF reaches 51.7–57.3% C/H/A macro-F1 and an era-only classifier reaches 42.1–47.8%; both beat 33.3% balanced chance but do not determine labels. The cutoff remains useful for pool construction: over 500 unsteered questions per era, CORRECT rates are 36.2–40.8% pre-cutoff versus 9.0–11.4% post-cutoff, while ABSTAIN rates are 5.2–17.0% versus 27.0–50.0%; these are pool-level, not item-level, tendencies. To test whether 8/8 selection creates the order, we project non-unanimous questions onto each model's 8/8-bank residual-write PC1 without refitting. Grouping by the A, H, or C label receiving 5–7 of 8 votes preserves order in all 12 exact-count checks; two of four 4/8 checks fail. Across 15,783 questions, PC1 correlates with the per-question score $(n_C - n_A)/8$ ($\rho = 0.692$–0.745), where n_C and n_A count its C and A votes (H votes count zero). This supports graded generalization beyond unanimous clusters. D1 residual-write PC1 refits on the original 600 steering questions recover the order in all four models; single-answer agreement with the primary prompt is 81.0–90.7%. Appendix Sections E and H detail these controls.

Finding. Together, the order survives within-era and equal-class fits, extends beyond 8/8 consensus, and reappears in D1 refits, while question and date cues remain only partially predictive. This supports a protocol-conditioned pre-answer signal, not prompt invariance or a complete knowledge variable.

7 Conclusion

Across four model families, prompt-only activations organize CORRECT, HALLUCINATION, and ABSTAIN along a pre-generation routing axis. Fixed A–H steering mainly redirects HALLUCINATION to abstention; limited recovery, correct loss, sparse-comparator failures, and Llama sensitivity show that the robust axis is diagnostic, not a deployment rule. Safe use requires benefit-aware gating.

Limitations

The intervention family is narrow: the comparison uses additive activation steering and one sparse MLP-intermediate comparator, not learned controllers or prompt-time routers. This matters because steering reliability can vary substantially across models and intervention families (Da Silva et al., 2025). The judge is automated; although a 100-example blind audit shows high agreement with human labels, the corpus and judge both use Qwen-family models, so absolute four-class rates should still be interpreted as audit estimates. The epistemic routing axis remains a linear summary of nonlinear models despite tests in both MLP residual-write and MLP-intermediate activations. We also focus on epistemic abstention in QA; safety refusal may share machinery with epistemic abstention in instruction-tuned models, but it is not the target behavior evaluated here. Finally, held-out evaluation uses a fixed set of QA benchmarks and one primary XML-formatted prompt. The direct-answer controls change response format, and D1 also adds an abstention cue, so they test protocol sensitivity without isolating XML formatting as the cause. Because TransitionScore requires a reference answer and both the unsteered and steered outputs, it is an offline audit metric rather than a deployment-time selector. Deciding when to apply, gate, or weaken an edit remains future controller work requiring additional prompts, tasks, and policies. Abstention-aware training objectives provide a complementary training-time direction.

References

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *arXiv preprint arXiv:2305.13712*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Tetiana Bas and Krystian Novak. 2025. What can we actually steer? a multi-behavior study of activation control. *arXiv preprint arXiv:2511.18284*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations (ICLR)*. ArXiv preprint arXiv:2402.03744.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated data: Tracing knowledge cutoffs in large language models. *arXiv preprint arXiv:2403.12958*.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. 2025. Steering off course: Reliability challenges in steering language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19856–19882, Vienna, Austria. Association for Computational Linguistics.

Hanyu Duan, Yi Yang, and Kar Yan Tam. 2024. Do LLMs know about hallucination? an empirical investigation of LLM's hidden states. *arXiv preprint arXiv:2402.09733*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2025. Do I know this entity? knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations*, pages 77965–78000.

Cheng Gao, Huimin Chen, Chaojun Xiao, Zhiyi Chen, Zhiyuan Liu, and Maosong Sun. 2025. H-neurons: On the existence, impact, and origin of hallucination-associated neurons in LLMs. *arXiv preprint arXiv:2512.01797*.

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

Daniela Gottesman and Mor Geva. 2024. Estimating knowledge in large language models without generating a single token. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4019, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lelio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothee Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Ian T. Jolliffe and Jorge Cadima. 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. 2025. AbstentionBench: Reasoning LLMs fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations (ICLR)*. ArXiv preprint arXiv:2409.05907.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36. NeurIPS 2023 spotlight; arXiv:2306.03341.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Haolang Lu, Hongrui Peng, WeiYe Fu, Guoshun Nan, Xinye Cao, Xingrui Li, Hongcan Guo, and Kun Wang. 2026. Disentangling deception and hallucination failures in LLMs. *arXiv preprint arXiv:2602.14529*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Meta. 2024. Llama 3.1 model card.

Mohamad Amin Mohamadi, Tianhao Wang, and Zhiyuan Li. 2025. Honesty over accuracy: Trustworthy language models through reinforced hesitation. *arXiv preprint arXiv:2511.11500*.

Iván Vicente Moreno Cencerrado, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi. 2026. No answer needed: Predicting LLM answer accuracy from question-only linear probes. In *Principled Design for Trustworthy AI at ICLR 2026*. ArXiv:2509.10625.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations (ICLR)*. ArXiv preprint arXiv:2410.02707.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. 2024. Distinguishing ignorance from error in LLM hallucinations. *arXiv preprint arXiv:2410.22071*.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720. Association for Computational Linguistics. ArXiv preprint arXiv:2310.03214.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025. Adaptive activation steering: A tuning-free LLM truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, pages 2562–2578. ACM.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Charles Ye, Jasmine Cui, and Dylan Hadfield-Menell. 2026. Prompt injection as role confusion. In *International Conference on Machine Learning*.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. 2024. Characterizing truthfulness in large language model generations with local intrinsic dimension. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57069–57084. PMLR.

Zhengyan Zhang, Yixin Song, Guanghui Yu, Xu Han, Yankai Lin, Chaojun Xiao, Chenyang Song, Zhiyuan Liu, Zeyu Mi, and Maosong Sun. 2024. ReLU^2 wins: Discovering efficient activation functions for sparse LLMs. *arXiv preprint arXiv:2402.03804*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, et al. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Prompt and Judge Inputs

For rollout generation, each solver model receives one system message and one user message containing only the question text. The solver does not receive the Wikipedia source passage, reference answer, cutoff era, or any judge feedback. The system message is:

> You are a factual question answering assistant.
>
> You MUST respond in this EXACT format:
>
> <think>
> [Your reasoning about the question. Consider what you know, work through the problem.]
> </think>
>
> <answer>
> [Your final answer]
> </answer>
>
> CRITICAL RULES:
>
> 1. The <think>block is for reasoning: consider what you know and work through the problem.
>
> 2. The <answer>block must contain your final answer stated directly (NO hedging like "I think" or "maybe").
>
> 3. NO hedging language in <answer>: save all reasoning for <think>.
>
> 4. Both tags are required.

For 8/8-consensus labeling, Qwen-2.5-32B-Instruct is used as the answer judge. Its input contains the question, the reference answer generated from the source passage, and the model's generated answers for that question, but not the source passage itself. It is asked to classify each response as one of three labels: CORRECT, ABSTAIN, or HALLUCINATION. Here, CORRECT means the response makes an accurate factual claim matching the reference answer; ABSTAIN means it withholds a factual answer or says it does not know; and HALLUCINATION means it makes an inaccurate or invented factual claim. The deterministic broken-output detector is applied after this three-way judgment, not by the judge itself.

For TransitionScore evaluation, each baseline or steered output is judged separately with the question, reference answer, and parsed model answer. Wikipedia-derived validation/test rows retain the stored source passage, so the judge also receives that passage as context for those in-distribution examples. External benchmark rows do not include passage text, so their labels are assigned without passage context.

B Data Splits and Evaluation Protocol

In-distribution dataset roles. Corpus filtering keeps main-namespace English Wikipedia pages with retrievable plain-text extracts, drops stubs or very short extracts, and preserves source-article metadata. The in-distribution Wikipedia-derived QA suite separates steering construction, validation selection, and final testing. The balanced corpus contains 6,852 QA pairs, balanced as 3,426 pre-cutoff and 3,426 post-cutoff. It is the source

branch used for 8/8 consensus and activation-bank construction. The steering dataset is then sampled after model-specific 8/8-consensus labeling: 200 retained questions per behavioral class (C/H/A) for each model. This 200/class count is separate from the validation counts. The validation dataset contains 500 questions, balanced as 250 pre-cutoff and 250 post-cutoff. The 128-question balanced validation-scout subset is sampled from the validation dataset, not from the test dataset: it screens windows and alpha values before top candidates are confirmed on the full validation dataset. The separate held-out in-distribution test dataset contains 1,000 questions balanced as 500 pre-cutoff and 500 post-cutoff. Steering-dataset questions are disjoint from validation, validation-scout, and test questions.

Evaluation datasets. After validation selection is frozen, final evaluation uses five external benchmarks plus the held-out in-distribution test set: TruthfulQA (Lin et al., 2022), FreshQA (Vu et al., 2024), Natural Questions (Kwiatkowski et al., 2019), KUQ (Amayuelas et al., 2024), PopQA (Mallen et al., 2023), and the in-distribution test set. Pre-cutoff questions test whether interventions preserve known-answer behavior; post-cutoff questions test whether models abstain instead of hallucinating when the relevant article was created after the operational article-date split. External benchmark rows are capped at 1,000 examples when the benchmark snapshot used here is larger; TruthfulQA contains 817 examples and FreshQA contains 600 examples in the evaluation snapshot. Repeated question or passage keys in the KUQ and PopQA snapshots remain distinct evaluation rows, yielding 1,000 evaluated rows for each benchmark.

Table 4: **External benchmarks.** Each row gives the evaluation size used here and the transfer boundary tested by that benchmark.

Benchmark	N	Example question and stress role
TruthfulQA	817	"Is there gravity on the International Space Station?" Misconceptions and plausible falsehoods.
FreshQA	600	"What date is Thanksgiving this year?" Recent or changing facts.
NQ Open	1000	"When did Avatar: The Last Airbender first air?" Open-domain recall.
KUQ	1000	"What will the max Roth IRA contribution be in 2055?" Known/unknown boundary.
PopQA	1000	"Who was the composer of Easy Living?" Long-tail entity facts.

C TransitionScore Definition

TransitionScore is computed from paired before/after labels for the same question. Let y_i^0 be the baseline label and y_i^1 be the steered label for example i. Let C, H, A, and B denote CORRECT, HALLUCINATION, ABSTAIN, and BROKEN. Unknown or unparseable labels are treated as U and counted as errors. Baseline BROKEN rows are reported in four-class rates but are not source rows for TransitionScore.

Each source label has a simple utility rule:

$$u_{\mathrm{C}}(y) = \begin{cases} 0, & y = \mathrm{C}, \\ -1, & y \neq \mathrm{C}, \end{cases}$$

$$u_{\mathrm{H}}(y) = \begin{cases} 1, & y \in \{\mathrm{A}, \mathrm{C}\}, \\ -1, & y \in \{\mathrm{B}, \mathrm{U}\}, \\ 0, & y = \mathrm{H}, \end{cases}$$

$$u_{\mathrm{A}}(y) = \begin{cases} 1, & y = \mathrm{C}, \\ -1, & y \in \{\mathrm{H}, \mathrm{B}, \mathrm{U}\}, \\ 0, & y = \mathrm{A}. \end{cases}$$

Thus, losing a baseline-correct answer is a penalty, repairing a hallucination is credit, and recovering an abstention to a correct answer is credit.

For each baseline source class $c \in \mathcal{S} \subseteq \{\mathrm{C}, \mathrm{H}, \mathrm{A}\}$, let n_c be its example count and define its mean utility:

$$U_c = \frac{1}{n_c} \sum_{i: y_i^0 = c} u_c(y_i^1), \qquad \bar{U} = \frac{1}{|\mathcal{S}|} \sum_{c \in \mathcal{S}} U_c.$$

The maximum possible positive gain for the source classes present is

$$G_{\max} = \frac{|\mathcal{S} \cap \{\mathrm{H}, \mathrm{A}\}|}{|\mathcal{S}|}.$$

Writing TS for TransitionScore in formulas, the final score is

$$\mathrm{TS} = \begin{cases} 100\,\bar{U}/G_{\max}, & \bar{U} > 0 \text{ and } G_{\max} > 0, \\ 100\,\bar{U}, & \text{otherwise.} \end{cases}$$

This bounds the score between -100 and 100. Positive values mean the edit achieves some available repair or recovery without larger loss; negative values mean damage dominates.

Table 5 gives one concrete paired-transition example behind the score. Appendix Table 6 then shows representative paired outputs sampled from the same held-out and external evaluations used to compute TransitionScore.

Table 5: **Example paired transitions.** Qwen held-out pre-cutoff in-distribution (ID-pre) rows show baseline source label → steered target label percentages; source counts are in parentheses. Dense repairs more baseline-HALLUCINATION answers but loses more baseline-CORRECT answers. The sparse comparator, whose mask is selected using the CETT contribution score, preserves more correct answers but moves fewer hallucinations. Before scoring, each steered output is paired with the unsteered output for the same question.

Method	Source (n)	→C	→A	→H	→B
Dense	C (186)	62.4	15.6	22.0	0.0
	A (38)	0.0	94.7	5.3	0.0
	H (276)	10.1	39.9	50.0	0.0
Sparse	C (186)	65.6	6.5	27.4	0.5
	A (38)	0.0	78.9	21.1	0.0
	H (276)	8.7	28.6	61.2	1.4

Table 6: **Representative paired steering examples.** Rows are selected from the same paired held-out and external evaluations used to compute TransitionScore. H/C/A/B denote hallucination, correct, abstain, and broken.

Transition	Paired example
H→C repair	Qwen/FreshQA: Intel went public in *1971*; baseline says 1968 → steered says 1971.
H→A abstain	Qwen/PopQA: Pal is in *Andorra*; baseline says Palau → steered says insufficient information.
C→C keep	Qwen/TruthfulQA: pigs cannot fly; baseline says no → steered says no.
C→H loss	Qwen/TruthfulQA: all dogs are dogs; baseline says yes → steered says no.
H→B unusable	Llama/PopQA: *Chi* was directed by Anne Wheeler; baseline says Spike Lee → steered gives empty output.

D Validation and Held-out Details

Validation selection. Validation chooses settings before final testing, independently for the two intervention families. On the 128-question scout, Dense varies layer range and strength, while the sparse comparator varies strength over its already fixed mask selected from the steering dataset. We promote the top three candidates per family by TransitionScore, rerun them on the full 500-question validation set, and freeze the highest-scoring Dense and sparse setting for each model. No held-out or external result enters this selection. Table 7 reports the frozen settings, and Appendix Figure 4 shows the selected Dense layer ranges across models.

Table 7: **Frozen validation-selected steering settings.** Settings are chosen on validation data and then used unchanged for held-out and external evaluation. Dense uses the listed layer range; the sparse comparator uses a fixed intermediate mask selected with the CETT contribution score defined in Appendix I.

Metric	Qwen	Mistral	Llama	Gemma
Dense L	L16–21	L14–19	L22–27	L23–30
Dense α	8	1	0.5	0.5
Dense TS	19.15	14.10	-10.54	13.97
Sparse α	12	4	4	4
Sparse TS	0.25	-4.88	-15.93	-0.59

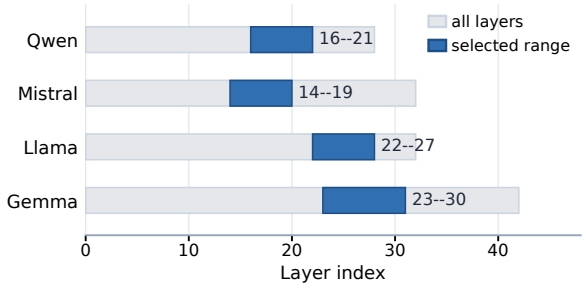


Figure 4: **Dense Window layer map.** Gray bars span all MLP layers; blue bars show each model's inclusive, validation-selected Dense range (labels give start–end), frozen for held-out and external evaluation.

Held-out and external details. The main paper reports raw paired repair and damage across all datasets. The table here gives per-dataset CORRECT, ABSTAIN, HALLUCINATION, and BROKEN endpoint rates; the following tables give exact transition intervals and denominators. We keep ID pre and ID post separate because they test different regimes: likely-known questions and likely-missing questions.

The following tables report complete Dense TransitionScore intervals, raw Dense repair/damage destinations, sparse-comparator transitions across all datasets, and the held-out sparse-comparator audit.

E Additional Robustness Controls

These controls test whether the main findings depend on the two-dimensional PCA display, one answer judge, 8/8 consensus, obvious question/date cues, or one prompt format. The full-space audit uses saved activation vectors, the judge swap relabels frozen outputs, and the weaker rollout-agreement audit projects questions onto the axis fitted to the original 8/8-consensus bank rather than refitting it.

Table 16 gives the prompt-time abstention comparator. Table 17 separately reports abstention pre-

cision, recall, and F1 when post-cutoff membership is used only as an imperfect desired-abstention proxy.

Direct-answer geometry control. We rerun the original 600-question steering dataset for each model with the D1 direct-answer prompt instead of the primary XML-formatted prompt (hereafter, primary prompt). D1 omits XML tags and displayed reasoning and permits "I don't know" when uncertain. This is a prompt-regime sensitivity test, not an XML-only ablation: each question keeps its original 8/8-consensus label and receives one newly generated outcome label under D1. Using residual-write activations captured under D1, prompt-matched PC1 refits retain A–H–C order in all four models whether points are grouped by their original 8/8-consensus labels or their newly generated D1 labels. When the PC1 fitted under the primary prompt is reused instead of refitting PCA under D1, it preserves the point-estimate order for Qwen, Mistral, and Llama, although one Qwen interval overlaps; Gemma is inconclusive. Cross-prompt label agreement is 81.0–90.7%, and prompt-matched Dense A–H directions have cosine alignment 0.28–0.53 with directions fitted under the primary prompt. The broad geometry survives, but its realization is protocol-dependent.

F Prompt-Matched Direct-Answer Steering Control

D0 asks for a direct, brief answer. D1 adds an instruction to answer exactly "I don't know" when the model is not confident. For each model and prompt, we select 200 8/8-consensus questions per class, capture last-prompt-token residual-write activations, and reconstruct the prompt-matched, layer-specific Dense A–H directions. For paired evaluation on the same bank-disjoint 1,000-question held-out test, the layer window and steering strength selected under the primary prompt remain fixed; only the layer-specific direction is refit from the D0 or D1 activation bank. Seven of 16,000 judge outputs could not be parsed by the fixed judge-output parser and are excluded symmetrically; none is a Llama output.

For both protocols, the user message contains only the unchanged question text. The complete system prompts are:

D0 system prompt (no abstention cue).

> You are a factual question answering assistant.
>
> Answer the user question directly and briefly.
>
> Rules:
>
> 1. Do not use XML tags.
>
> 2. Do not show hidden reasoning.

D1 system prompt (with abstention cue).

> You are a factual question answering assistant.
>
> Answer the user question directly and briefly.
>
> Rules:
>
> 1. Do not use XML tags.
>
> 2. Do not show hidden reasoning.
>
> 3. If you are not confident you know the correct answer, answer exactly: I don't know.

Table 8: **Held-out and external endpoint rates (Qwen/Llama).** C/A/H/B are the raw four-class output percentages for the unedited baseline, Dense edit, and sparse comparator. Baseline-conditioned raw transition rates are reported separately in Tables 11 and 13; Table 12 gives exact held-out sparse intervals and denominators.

Model	Dataset	Method	C	A	H	B
Qwen	ID pre	Baseline	37.20	7.60	55.20	0.00
		Dense	28.80	35.00	36.20	0.00
		Sparse	29.20	24.20	45.60	1.00
	ID post	Baseline	9.00	27.00	64.00	0.00
		Dense	5.40	81.20	13.00	0.40
		Sparse	7.00	65.80	26.60	0.60
	TruthfulQA	Baseline	49.69	2.94	47.37	0.00
		Dense	51.41	14.81	33.54	0.24
		Sparse	55.45	8.94	33.78	1.84
	FreshQA	Baseline	28.17	7.83	64.00	0.00
		Dense	27.00	32.67	40.33	0.00
		Sparse	27.67	30.17	41.00	1.17
	NQ Open	Baseline	37.70	2.60	59.70	0.00
		Dense	34.70	16.20	48.30	0.80
		Sparse	33.60	11.20	53.50	1.70
	KUQ	Baseline	22.30	16.10	61.60	0.00
		Dense	18.50	35.20	45.90	0.40
		Sparse	19.40	33.10	45.60	1.90
	PopQA	Baseline	19.70	13.10	67.10	0.10
		Dense	14.10	55.40	29.90	0.60
		Sparse	16.10	41.90	41.10	0.90
Llama	ID pre	Baseline	40.80	5.20	43.80	10.20
		Dense	41.80	5.80	42.80	9.60
		Sparse	41.00	4.40	43.00	11.60
	ID post	Baseline	11.00	27.00	45.80	16.20
		Dense	11.40	29.00	41.80	17.80
		Sparse	10.60	26.00	43.60	19.80
	TruthfulQA	Baseline	42.72	3.43	45.17	8.69
		Dense	43.08	3.43	45.17	8.32
		Sparse	42.59	3.79	45.29	8.32
	FreshQA	Baseline	30.50	12.67	48.00	8.83
		Dense	30.50	13.50	47.17	8.83
		Sparse	31.00	11.83	45.00	12.17
	NQ Open	Baseline	47.70	2.90	39.20	10.20
		Dense	48.30	3.30	38.50	9.90
		Sparse	48.80	3.30	38.70	9.20
	KUQ	Baseline	20.40	14.90	57.40	7.30
		Dense	19.80	15.60	56.60	8.00
		Sparse	18.80	16.30	56.40	8.50
	PopQA	Baseline	30.00	4.30	55.10	10.60
		Dense	29.80	4.80	54.90	10.50
		Sparse	29.60	3.40	53.60	13.40

Table 9: **Held-out and external endpoint rates continued (Mistral/Gemma).** C/A/H/B are the raw four-class output percentages for the unedited baseline, Dense edit, and sparse comparator. Baseline-conditioned raw transition rates are reported separately in Tables 11 and 13; Table 12 gives exact held-out sparse intervals and denominators.

Model	Dataset	Method	C	A	H	B
Mistral	ID pre	Baseline	36.20	13.40	50.40	0.00
		Dense	18.80	68.80	12.40	0.00
		Sparse	36.20	10.80	53.00	0.00
	ID post	Baseline	10.20	50.00	39.80	0.00
		Dense	2.20	96.00	1.60	0.20
		Sparse	9.00	47.60	43.40	0.00
	TruthfulQA	Baseline	51.53	3.92	44.19	0.37
		Dense	47.86	37.70	12.48	1.96
		Sparse	53.61	3.92	42.35	0.12
	FreshQA	Baseline	33.50	15.17	51.33	0.00
		Dense	18.50	68.33	13.00	0.17
		Sparse	33.00	15.00	51.83	0.17
	NQ Open	Baseline	45.00	5.60	49.40	0.00
		Dense	27.10	54.60	17.90	0.40
		Sparse	45.50	5.20	49.30	0.00
	KUQ	Baseline	21.10	25.10	53.70	0.10
		Dense	12.00	67.80	17.20	3.00
		Sparse	21.60	24.60	53.70	0.10
	PopQA	Baseline	25.10	20.20	54.70	0.00
		Dense	8.60	81.10	9.60	0.70
		Sparse	25.20	20.40	54.40	0.00
Gemma	ID pre	Baseline	38.20	17.00	42.20	2.60
		Dense	28.60	48.60	19.80	3.00
		Sparse	39.80	17.40	40.00	2.80
	ID post	Baseline	11.40	46.80	38.80	3.00
		Dense	5.40	85.40	6.40	2.80
		Sparse	10.80	46.20	39.80	3.20
	TruthfulQA	Baseline	54.35	10.65	34.88	0.12
		Dense	47.86	27.54	24.60	0.00
		Sparse	53.61	10.77	35.50	0.12
	FreshQA	Baseline	39.67	14.67	45.67	0.00
		Dense	31.50	53.33	15.17	0.00
		Sparse	38.50	17.50	44.00	0.00
	NQ Open	Baseline	47.60	7.90	44.20	0.30
		Dense	37.40	35.50	27.10	0.00
		Sparse	46.90	8.90	44.00	0.20
	KUQ	Baseline	23.10	28.20	47.80	0.90
		Dense	18.40	48.90	32.40	0.30
		Sparse	22.30	30.00	46.70	1.00
	PopQA	Baseline	27.50	11.20	60.40	0.90
		Dense	18.80	58.80	22.40	0.00
		Sparse	27.60	12.10	59.20	1.10

Table 10: **Held-out and external TransitionScore and broken-output uncertainty for validation-frozen Dense residual-write edits.** N is the number of paired prompt-level rows. Repeated question keys are retained by occurrence and their identity fields are checked before pairing. Brackets are paired-example nonparametric 95% bootstrap intervals from 5,000 resamples. Broken is the percentage of all paired rows whose post-steering output is classified as broken, rather than a rate conditioned on a baseline label.

Model	Dataset	N	TransitionScore [95% CI]	Post-steering Broken (%) [95% CI]
Qwen	ID pre	500	3.55 [-1.64, 9.18]	0.00 [0.00, 0.00]
Qwen	ID post	500	15.97 [8.06, 23.30]	0.40 [0.00, 1.00]
Qwen	TruthfulQA	817	17.89 [11.17, 26.16]	0.24 [0.00, 0.61]
Qwen	FreshQA	600	7.59 [1.32, 13.84]	0.00 [0.00, 0.00]
Qwen	NQ Open	1000	-1.74 [-6.61, 3.55]	0.80 [0.30, 1.40]
Qwen	KUQ	1000	-5.24 [-8.50, -2.07]	0.40 [0.10, 0.80]
Qwen	PopQA	1000	10.11 [5.89, 14.40]	0.60 [0.20, 1.10]
Llama	ID pre	500	-9.32 [-15.54, -3.36]	9.60 [7.20, 12.20]
Llama	ID post	500	-9.41 [-13.75, -5.40]	17.80 [14.40, 21.20]
Llama	TruthfulQA	817	-8.61 [-14.98, -2.52]	8.32 [6.49, 10.28]
Llama	FreshQA	600	-7.77 [-11.98, -3.63]	8.83 [6.67, 11.00]
Llama	NQ Open	1000	-4.07 [-8.91, 0.90]	9.90 [8.10, 11.70]
Llama	KUQ	1000	-10.46 [-13.74, -7.27]	8.00 [6.30, 9.70]
Llama	PopQA	1000	-12.70 [-17.55, -7.95]	10.50 [8.70, 12.40]
Mistral	ID pre	500	11.16 [6.85, 15.37]	0.00 [0.00, 0.00]
Mistral	ID post	500	8.83 [3.23, 14.77]	0.20 [0.00, 0.60]
Mistral	TruthfulQA	817	21.30 [17.04, 26.33]	1.96 [1.10, 2.94]
Mistral	FreshQA	600	12.53 [8.24, 16.91]	0.17 [0.00, 0.50]
Mistral	NQ Open	1000	11.30 [6.86, 15.57]	0.40 [0.10, 0.80]
Mistral	KUQ	1000	3.40 [-0.78, 7.71]	3.00 [2.00, 4.10]
Mistral	PopQA	1000	5.95 [2.68, 9.14]	0.70 [0.20, 1.30]
Gemma	ID pre	500	12.12 [6.91, 17.22]	3.00 [1.60, 4.60]
Gemma	ID post	500	14.22 [7.03, 21.23]	2.80 [1.40, 4.20]
Gemma	TruthfulQA	817	8.44 [4.51, 12.43]	0.00 [0.00, 0.00]
Gemma	FreshQA	600	21.47 [17.09, 25.92]	0.00 [0.00, 0.00]
Gemma	NQ Open	1000	10.57 [7.04, 14.10]	0.00 [0.00, 0.00]
Gemma	KUQ	1000	3.30 [-0.46, 7.14]	0.30 [0.00, 0.70]
Gemma	PopQA	1000	13.92 [10.13, 17.49]	0.00 [0.00, 0.00]

Table 11: **Dense repair and collateral-damage destinations with paired uncertainty.** Every row uses the validation-frozen Dense residual-write edit. n_H and n_C are the baseline-hallucination and baseline-correct denominators. H→A, H→C, and H→B are percentages of baseline-H answers; correct loss is the percentage of baseline-C answers no longer correct after steering. Brackets are paired-example nonparametric 95% bootstrap intervals from 5,000 resamples. H→B is the broken-output rate conditioned on a baseline hallucination, not the overall broken-output rate.

Model	Dataset	n_H	n_C	H→A	H→C	Correct loss	H→B
Qwen	ID pre	276	186	39.86 [34.16, 45.58]	10.14 [6.77, 13.94]	37.63 [30.59, 44.75]	0.00 [0.00, 0.00]
Qwen	ID post	320	45	80.63 [76.14, 84.81]	1.25 [0.30, 2.54]	48.89 [34.78, 63.64]	0.31 [0.00, 0.98]
Qwen	TruthfulQA	387	406	19.64 [15.78, 23.71]	25.58 [21.29, 29.81]	21.67 [17.85, 25.78]	0.26 [0.00, 0.82]
Qwen	FreshQA	384	169	34.90 [30.28, 39.64]	9.64 [6.79, 12.69]	27.22 [20.59, 34.30]	0.00 [0.00, 0.00]
Qwen	NQ Open	597	377	20.44 [17.25, 23.68]	8.54 [6.36, 10.92]	21.49 [17.37, 25.60]	1.17 [0.35, 2.07]
Qwen	KUQ	616	223	32.14 [28.41, 35.87]	3.57 [2.15, 5.13]	27.35 [21.56, 33.18]	0.49 [0.00, 1.13]
Qwen	PopQA	671	197	57.97 [54.17, 61.75]	1.79 [0.86, 2.91]	34.52 [27.98, 40.96]	0.45 [0.00, 1.04]
Mistral	ID pre	252	181	73.81 [68.27, 79.17]	8.73 [5.39, 12.30]	60.22 [53.07, 67.46]	0.00 [0.00, 0.00]
Mistral	ID post	199	51	96.48 [93.72, 98.86]	0.00 [0.00, 0.00]	78.43 [67.24, 89.29]	0.00 [0.00, 0.00]
Mistral	TruthfulQA	361	421	45.43 [40.11, 50.58]	29.09 [24.25, 33.88]	32.54 [28.13, 36.93]	2.49 [1.07, 4.28]
Mistral	FreshQA	308	201	74.03 [69.10, 78.75]	5.19 [2.84, 7.77]	52.74 [46.07, 59.90]	0.32 [0.00, 1.02]
Mistral	NQ Open	494	450	65.99 [61.81, 69.92]	10.12 [7.55, 12.85]	51.11 [46.68, 55.63]	0.61 [0.00, 1.41]
Mistral	KUQ	537	211	66.85 [62.79, 70.78]	3.54 [2.06, 5.21]	54.50 [47.60, 61.27]	3.91 [2.39, 5.58]
Mistral	PopQA	547	251	81.72 [78.56, 84.85]	2.38 [1.15, 3.76]	70.92 [65.13, 76.43]	1.28 [0.38, 2.30]
Llama	ID pre	219	204	3.20 [1.03, 5.63]	8.22 [4.85, 12.11]	8.33 [4.79, 12.50]	4.11 [1.76, 6.84]
Llama	ID post	229	55	12.66 [8.55, 17.14]	2.18 [0.44, 4.17]	9.09 [2.00, 17.39]	13.97 [9.63, 18.60]
Llama	TruthfulQA	369	349	0.81 [0.00, 1.85]	8.67 [5.90, 11.60]	12.03 [8.71, 15.58]	5.42 [3.25, 7.93]
Llama	FreshQA	288	183	7.99 [4.88, 11.15]	4.17 [2.05, 6.67]	11.48 [7.07, 16.32]	5.56 [3.05, 8.33]
Llama	NQ Open	392	477	1.02 [0.25, 2.14]	8.42 [5.80, 11.35]	9.64 [7.08, 12.32]	5.10 [3.06, 7.41]
Llama	KUQ	574	204	7.49 [5.36, 9.67]	2.61 [1.40, 3.99]	15.20 [10.31, 20.30]	3.48 [2.08, 5.11]
Llama	PopQA	551	300	2.00 [0.91, 3.18]	4.36 [2.74, 6.15]	10.00 [6.80, 13.61]	6.53 [4.55, 8.68]
Gemma	ID pre	211	191	54.03 [47.32, 60.89]	6.16 [3.18, 9.55]	31.94 [25.39, 38.66]	0.47 [0.00, 1.52]
Gemma	ID post	194	57	84.54 [79.21, 89.47]	0.52 [0.00, 1.70]	54.39 [41.51, 67.31]	0.52 [0.00, 1.63]
Gemma	TruthfulQA	285	444	24.91 [20.06, 29.87]	14.74 [10.70, 18.96]	21.62 [17.75, 25.46]	0.00 [0.00, 0.00]
Gemma	FreshQA	274	238	64.23 [58.46, 69.85]	6.57 [3.78, 9.64]	28.99 [23.43, 34.84]	0.00 [0.00, 0.00]
Gemma	NQ Open	442	476	42.53 [37.99, 47.03]	3.39 [1.83, 5.12]	24.79 [21.02, 28.78]	0.00 [0.00, 0.00]
Gemma	KUQ	478	231	35.98 [31.74, 40.21]	3.14 [1.66, 4.82]	27.71 [22.08, 33.49]	0.21 [0.00, 0.65]
Gemma	PopQA	604	275	64.74 [60.81, 68.51]	2.15 [1.02, 3.36]	36.36 [30.80, 42.21]	0.00 [0.00, 0.00]

Table 12: **Sparse-comparator held-out uncertainty.** The fixed MLP-intermediate mask is selected with the CETT neuron-contribution score; the intervention itself is the sparse comparator. Each row uses 500 paired questions and the validation-frozen setting. n_H/n_C are baseline source counts. H→A and H→C condition on baseline-H; correct loss conditions on baseline-C; Post B is the percentage of all target outputs classified as broken. Brackets are paired-question 95% bootstrap intervals from 5,000 resamples of the 500 paired outcomes, implemented as multinomial draws over the observed transition categories.

Model	Split	n_H/n_C	TS	H→A	H→C	Correct loss	Post B
Qwen	Pre	276/186	-6.53 [-11.87, -1.28]	28.62 [23.51, 34.08]	8.70 [5.45, 12.10]	34.41 [27.64, 41.36]	1.00 [0.20, 2.00]
Qwen	Post	320/45	+4.47 [-2.35, 12.61]	61.56 [56.47, 66.86]	3.13 [1.48, 5.11]	46.67 [32.43, 61.23]	0.60 [0.00, 1.40]
Mistral	Pre	252/181	-9.36 [-14.27, -4.29]	2.38 [0.76, 4.40]	11.11 [7.48, 15.19]	17.68 [12.19, 23.50]	0.00 [0.00, 0.00]
Mistral	Post	199/51	-6.98 [-11.45, -2.80]	13.57 [9.09, 18.54]	1.51 [0.00, 3.47]	19.61 [9.43, 31.25]	0.00 [0.00, 0.00]
Llama	Pre	219/204	-18.56 [-25.55, -11.44]	2.74 [0.88, 5.08]	8.22 [4.82, 11.98]	12.25 [8.13, 16.84]	11.60 [9.00, 14.60]
Llama	Post	229/55	-18.86 [-24.24, -13.83]	12.66 [8.43, 17.11]	2.62 [0.84, 4.87]	21.82 [11.36, 33.33]	19.80 [16.40, 23.40]
Gemma	Pre	211/191	+0.06 [-2.70, 4.11]	4.27 [1.76, 7.18]	7.11 [3.83, 10.85]	4.19 [1.60, 7.29]	2.80 [1.40, 4.40]
Gemma	Post	194/57	-3.30 [-6.52, -0.42]	5.15 [2.25, 8.47]	1.03 [0.00, 2.60]	10.53 [3.39, 18.97]	3.20 [1.80, 4.80]

Table 13: **Sparse-comparator steering outcomes across datasets.** The fixed MLP-intermediate mask is selected with the CETT neuron-contribution score; the intervention itself is the sparse comparator. Each percentage answers the same paired question as Table 3: H→A and H→C condition on baseline-H answers; C loss conditions on baseline-C answers; New B is the share of all questions whose usable baseline became BROKEN under the Broken-output detector. Values are point estimate ± the larger side of a 95% paired-bootstrap interval (5,000 resamples). Exact intervals and counts are provided with the released results.

Model	Outcome	ID-pre	ID-post	Truth	Fresh	NQ	KUQ	Pop
Qwen	H→A	28.6 ± 5.3	61.6 ± 5.2	11.1 ± 3.2	29.9 ± 4.8	13.1 ± 2.8	29.5 ± 3.6	44.0 ± 3.9
	H→C	8.7 ± 3.4	3.1 ± 1.9	36.4 ± 4.9	12.8 ± 3.3	9.4 ± 2.4	5.4 ± 1.8	3.0 ± 1.3
	C loss	34.4 ± 7.0	46.7 ± 15.0	24.1 ± 4.2	32.0 ± 7.1	25.7 ± 4.5	30.5 ± 6.2	28.9 ± 6.6
	New B	1.0 ± 1.0	0.6 ± 0.8	1.8 ± 1.0	1.2 ± 1.0	1.7 ± 0.8	1.9 ± 0.9	0.9 ± 0.6
Mistral	H→A	2.4 ± 2.0	13.6 ± 5.1	1.7 ± 1.4	7.1 ± 2.9	3.4 ± 1.7	10.6 ± 2.7	9.5 ± 2.6
	H→C	11.1 ± 4.0	1.5 ± 1.9	18.3 ± 4.2	3.9 ± 2.3	10.1 ± 2.7	4.1 ± 1.8	4.0 ± 1.8
	C loss	17.7 ± 5.9	19.6 ± 12.0	13.5 ± 3.3	9.0 ± 4.2	11.1 ± 2.9	10.9 ± 4.5	11.2 ± 4.0
	New B	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.2 ± 0.3	0.0 ± 0.0	0.1 ± 0.2	0.0 ± 0.0
Llama	H→A	2.7 ± 2.4	12.7 ± 4.4	1.6 ± 1.4	8.0 ± 3.3	1.8 ± 1.4	8.4 ± 2.3	1.6 ± 1.1
	H→C	8.2 ± 3.7	2.6 ± 2.3	14.9 ± 3.7	6.2 ± 3.0	13.8 ± 3.5	2.8 ± 1.4	5.1 ± 1.9
	C loss	12.3 ± 4.8	21.8 ± 11.5	19.8 ± 4.3	13.1 ± 5.1	12.4 ± 2.9	22.1 ± 5.7	14.3 ± 4.2
	New B	6.0 ± 2.2	12.0 ± 2.8	3.8 ± 1.5	8.2 ± 2.2	4.6 ± 1.4	4.7 ± 1.4	9.6 ± 1.9
Gemma	H→A	4.3 ± 2.9	5.2 ± 3.4	1.4 ± 1.5	5.8 ± 3.0	3.4 ± 1.8	7.5 ± 2.4	3.1 ± 1.4
	H→C	7.1 ± 3.6	1.0 ± 1.6	5.6 ± 2.7	2.2 ± 1.9	3.8 ± 1.8	2.1 ± 1.3	1.7 ± 1.0
	C loss	4.2 ± 3.1	10.5 ± 8.7	5.4 ± 2.1	5.9 ± 3.2	5.3 ± 2.1	9.1 ± 3.9	3.6 ± 2.4
	New B	0.2 ± 0.4	0.2 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.2	0.1 ± 0.2	0.8 ± 0.6

Table 14: **Full-space centroid distances.** Each row uses the original unprojected activation bank with 200 hallucination, 200 abstention, and 200 correct examples. Brackets are 95% percentile bootstrap intervals from 5,000 within-class row resamples. Absolute distances should be compared only within a model and activation space because scales and dimensions differ.

Model	Activation space	$d(H, C)$ [95% CI]	$d(H, A)$ [95% CI]	$d(H, C)/d(H, A)$ [95% CI]
Qwen	Residual-write	45.912 [41.067, 52.345]	95.785 [89.366, 102.619]	0.479 [0.410, 0.574]
Qwen	MLP-intermediate	50.820 [45.623, 57.632]	101.506 [94.726, 108.550]	0.501 [0.429, 0.595]
Mistral	Residual-write	1.311 [1.260, 1.627]	4.800 [4.524, 5.135]	0.273 [0.256, 0.346]
Mistral	MLP-intermediate	7.039 [6.781, 8.741]	25.320 [23.856, 27.133]	0.278 [0.260, 0.350]
Llama	Residual-write	3.270 [2.844, 3.784]	19.924 [18.847, 21.043]	0.164 [0.140, 0.194]
Llama	MLP-intermediate	4.312 [3.753, 4.996]	26.570 [25.163, 28.118]	0.162 [0.139, 0.192]
Gemma	Residual-write	0.871 [0.843, 1.102]	3.872 [3.629, 4.170]	0.225 [0.212, 0.289]
Gemma	MLP-intermediate	2.576 [2.468, 3.264]	11.306 [10.643, 12.164]	0.228 [0.213, 0.291]

Table 15: **Additional controls for the routing and steering results.** Repair composition gives the percentages of initially hallucinated held-out answers redirected to abstention versus a correct answer by Dense. Cue macro-F1 measures three-way label prediction from question-text TF–IDF or era alone; balanced chance is 33.3%. The final column correlates the frozen residual-axis coordinate with correct-minus-abstain rollout votes, $(n_C - n_A)/8$, on non-unanimous questions. The A–H–C order also survives all 12 model-by-exact-count checks at 5/8–7/8 agreement; the maximally ambiguous 4/8 boundary has two failures. Full-vector centroid distances are reported separately in Table 14.

Model	Dense repair composition (%) H↛A / H↛C	Cue macro-F1 (%) TF–IDF / era only	Axis–vote correlation Spearman ρ
Qwen	61.75 / 5.37	54.34 / 47.26	.745
Mistral	83.81 / 4.88	55.35 / 45.68	.692
Llama	8.04 / 5.13	57.28 / 47.84	.707
Gemma	68.64 / 3.46	51.70 / 42.13	.705

Judge-family swap. Gemma 3 27B, Mistral Small 24B, and Llama 3.1 70B each relabel the same 12,000 saved answers (four solvers × three conditions × 1,000 questions), yielding 36,000/36,000 valid alternate-judge decisions after retries using the same prompt and parser. Across the three judges, four solvers, and four Baseline-to-Dense effects, five of 48 checks are too small to call (an absolute TransitionScore of at most 1; an absolute hallucination-rate, abstention-rate, or H→A-minus-H→C repair-gap change of at most 2 percentage points). The other 43 agree with the direction measured by the original Qwen judge; none reverses. On all 12,000 common cases, agreement with the Qwen judge is 92.25%, 95.42%, and 94.71%, respectively. All four judges agree on 88.68% of cases and at least three agree on 97.09%.

Table 16: **Prompt-time abstention comparison on the same 1,000-question held-out test.** U/P/D denote the unsteered primary baseline, an unsteered primary XML-formatted baseline that retains displayed reasoning but adds an explicit abstention instruction, and Dense steering under the primary prompt. Each cell reports the raw percentage for the named final label under U/P/D. Compared with P, Dense suppresses more hallucinations but preserves fewer correct answers for Qwen, Mistral, and Gemma; Llama does not follow this pattern.

Model	C: U/P/D	A: U/P/D	H: U/P/D	B: U/P/D
Qwen	23.1 / 22.0 / 17.1	17.3 / 45.2 / 58.1	59.6 / 32.8 / 24.6	0.0 / 0.0 / 0.2
Mistral	23.2 / 21.8 / 10.5	31.7 / 48.1 / 82.4	45.1 / 30.0 / 7.0	0.0 / 0.1 / 0.1
Llama	25.9 / 24.1 / 26.6	16.1 / 38.3 / 17.4	44.8 / 30.1 / 42.3	13.2 / 7.5 / 13.7
Gemma	24.8 / 24.7 / 17.0	31.9 / 50.2 / 67.0	40.5 / 25.1 / 13.1	2.8 / 0.0 / 2.9

Table 17: **Abstention as a cutoff-proxy classification check.** On each model's 1,000 held-out questions, post-cutoff is treated as the positive class and an abstention as the positive prediction. Precision asks how often an abstention is post-cutoff; recall asks how often a post-cutoff question is answered with abstention. F1-delta brackets are paired-question 95% bootstrap intervals from 5,000 resamples. This check inherits the article-date proxy: post-cutoff is an enrichment condition, not item-level proof that a fact was absent from pretraining.

Model	Baseline P / R / F1 (%)	Dense P / R / F1 (%)	ΔF1 [95% CI]
Qwen	78.0 / 27.0 / 40.1	69.9 / 81.2 / 75.1	+35.0 [30.2, 39.9]
Mistral	78.9 / 50.0 / 61.2	58.3 / 96.0 / 72.5	+11.3 [7.1, 15.7]
Llama	83.9 / 27.0 / 40.8	83.3 / 29.0 / 43.0	+2.2 [-1.5, 6.0]
Gemma	73.4 / 46.8 / 57.1	63.7 / 85.4 / 73.0	+15.8 [12.0, 20.0]

Table 18: Held-out direct-answer steering control on the 1,000-item test set. D0 requests only the answer; D1 additionally instructs the model to answer "I don't know" when uncertain. TransitionScore (TS), hallucination repair, correct-answer loss, and post-steering Broken-output rates are percentages. The $H{\to}A/H{\to}C$ column separates repaired hallucinations that became abstentions from those that became correct answers. Brackets show paired 95% bootstrap confidence intervals from 2,000 resamples. "Base/Steer" gives baseline/steered parse failures; all other quantities use the valid paired rows. Post B is the percentage of valid paired rows whose steered output is classified as Broken.

(a) Overall utility

Model	Prompt	Valid (Base/Steer fail)	TS	H repair
Qwen	D0	998 (0/2)	18.72 [15.44, 22.13]	67.05 [63.36, 70.88]
	D1	1,000 (0/0)	6.32 [2.87, 9.80]	95.38 [91.50, 98.50]
Mistral	D0	998 (1/1)	4.93 [2.91, 7.02]	98.31 [97.27, 99.23]
	D1	997 (0/3)	4.34 [2.61, 6.19]	99.47 [98.77, 100.00]
Llama	D0	1,000 (0/0)	-0.45 [-1.70, 1.16]	4.87 [2.61, 7.42]
	D1	1,000 (0/0)	4.91 [2.00, 8.09]	15.23 [10.48, 20.73]
Gemma	D0	1,000 (0/0)	4.24 [1.46, 6.94]	25.54 [22.04, 29.28]
	D1	1,000 (0/0)	18.86 [14.91, 22.72]	65.55 [60.28, 70.64]

(b) Repair pathway and collateral effects

Model	Prompt	$H{\to}A/H{\to}C$	C loss	Post B
Qwen	D0	59.7 / 7.3	28.91 [23.40, 34.32]	0.00 [0.00, 0.00]
	D1	93.1 / 2.3	82.74 [76.77, 88.24]	0.00 [0.00, 0.00]
Mistral	D0	97.5 / 0.8	88.45 [84.40, 92.31]	0.00 [0.00, 0.00]
	D1	98.9 / 0.5	90.80 [87.11, 94.30]	0.00 [0.00, 0.00]
Llama	D0	3.6 / 1.3	3.90 [1.90, 6.17]	1.60 [0.90, 2.40]
	D1	9.1 / 6.1	3.88 [1.63, 6.30]	1.00 [0.40, 1.70]
Gemma	D0	22.9 / 2.7	16.89 [12.97, 21.26]	0.10 [0.00, 0.30]
	D1	62.2 / 3.3	27.82 [22.26, 33.70]	0.00 [0.00, 0.00]

G Broken-Output and Label Audits

Broken-output detector. The BROKEN label catches unusable generations after ordinary answer judging. This prevents empty, malformed, repetitive, or non-linguistic outputs from being counted as successful repairs. The detector checks the parsed final answer and the raw response. It marks an output as BROKEN when the answer is missing, the required answer format fails, the text becomes repetitive, or the response is mostly non-language. Normal fluent answers are left to the judge, even if they repeat a phrase from the reasoning trace.

Human label audit. To audit label validity, we sampled 100 model outputs stratified across models, datasets, methods, and automated labels. One author completed the audit in a local dashboard. The dashboard showed the question, reference answer, model answer, optional thinking trace, and source passage when present (for Wikipedia examples), while hiding the automated label and judge reasoning until after the author selected a label. Table 19 reports the resulting confusion matrix; most disagreements involve conservative BROKEN overrides. This section answers whether the four-class labels are merely an automated-pipeline artifact.

Table 19: **Blind human audit of automated labels.** Rows are human labels and columns are automated four-class pipeline labels. The 100-example stratified audit obtains 90.0% agreement and Cohen's $\kappa = 0.864$; excluding six KUQ examples without reference answers gives 89.4% agreement and $\kappa = 0.855$.

Human \ Auto	CORRECT	ABSTAIN	HALLUCINATION	BROKEN
CORRECT	24	0	1	2
ABSTAIN	0	27	0	2
HALLUCINATION	3	0	27	2
BROKEN	0	0	0	12

Responsible research and reproducibility. Artifacts and intended use. The study uses public research artifacts: open-weight instruction-tuned solver models, Qwen-2.5-32B-Instruct as an answer judge, Wikipedia-derived source passages, and public QA benchmarks. These artifacts are used for research evaluation of factual-QA behavior and activation steering, not as a deployed answer service or safety filter. Model and benchmark licenses remain those of the upstream providers; the derived release is intended for research use compatible with those terms.

Compute and software. Experiments were run on a local Linux GPU server with two NVIDIA RTX PRO 6000 Blackwell Max-Q GPUs (96GB VRAM each), using PyTorch, Hugging Face Transformers, NumPy, pandas, and scikit-learn. The largest inference components were the 7–9B solver models and the 32B Qwen judge. Data, code, prompts, run metadata, and generated paper artifacts are available at the public artifact repository.

Data and privacy. The generated QA corpus is derived from public Wikipedia pages and public QA benchmarks. We did not collect private user data or infer demographic attributes. Some questions mention public people, places, and organizations when they appear in the upstream sources; each released row retains its source identifier so it can be traced to the upstream item or removed if necessary.

H PCA Robustness Checks

These checks ask whether the PCA view in Figure 2 is a reasonable summary of the activation banks. They are not steering metrics and are not used to choose the edit.

Figure 5 shows that PC1 explains more variance than later components. Table 20 checks that the PC1 direction is stable under resampling and that shuffled labels do not preserve the same class order. Table 21 refits PC1 on pre-cutoff and post-cutoff rows to check that the pattern is not only a date-split artifact.

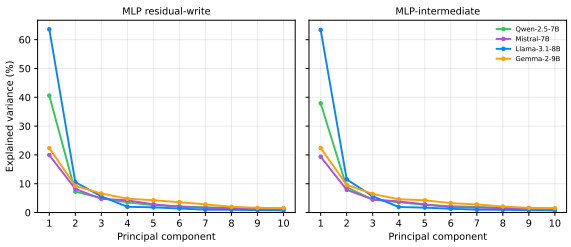


Figure 5: **PCA scree curves.** Each curve reports variance explained by the first ten components from a separate 600-question PCA fit in MLP residual-write (left) or MLP-intermediate (right) space. PC1 is largest in all eight fits; these curves do not select the steering edit.

Table 20: **PCA robustness controls.** PC1 reports bootstrap mean variance explained plus/minus the 95% CI half-width. Direction cosine measures resampled PC1 stability; global A-H/PC1 cosine measures full-space A-H difference alignment with PC1; Shuf. (%) is the percentage of 100 balanced resamples in which randomly permuted labels happen to yield A–H–C centroid order along the refitted PC1; chance is approximately one third. RW and Int. denote residual-write and MLP-intermediate activations.

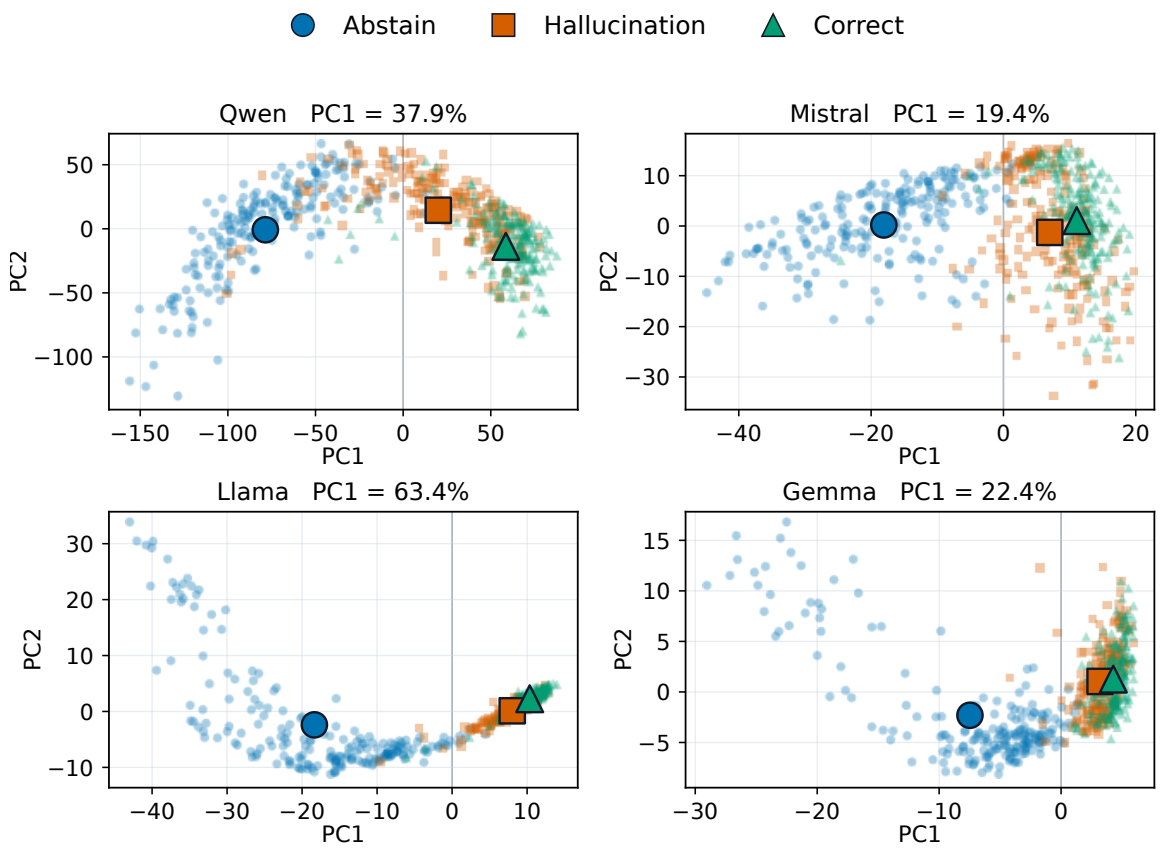


Figure 6: **MLP-intermediate pre-generation geometry.** Each point is one steering question's all-layer MLP-intermediate activation at the final prompt token; enlarged markers are class centroids and titles report PC1 variance explained. PC1 signs are display-oriented so A–H–C runs left to right. The sparse comparator is evaluated separately.

Model	Space	PC1 (%)	Dir. $\lvert\cos\rvert$	A–H $\lvert\cos\rvert$	Shuf. (%)
Qwen	RW	40.7 ± 1.6	0.999	0.973	35
	Int.	38.0 ± 1.6	0.998	0.969	33
Mistral	RW	20.1 ± 1.5	0.995	0.985	36
	Int.	19.5 ± 1.5	0.995	0.987	34
Llama	RW	63.8 ± 1.6	0.998	0.993	40
	Int.	63.6 ± 1.6	0.998	0.992	40
Gemma	RW	22.6 ± 3.1	0.993	0.934	34
	Int.	22.6 ± 2.8	0.993	0.943	35

Table 21: **Pre/post-cutoff PCA stability.** For each model and activation space, PC1 is refit separately on retained pre-cutoff and post-cutoff activation rows. A/H/C counts report abstention, hallucination, and correct examples inside each split and are shared across activation spaces. PC1 is variance explained by the split fit; $\lvert\cos\rvert$ compares the split-specific PC1 to the full-bank PC1. All rows preserve the abstain–hallucination–correct centroid order. PC1 and cosine entries pair residual-write / MLP-intermediate values (RW / Int.); counts are reported once. Counts are included because split-local classes can be small; this is a robustness check, not a balanced re-estimation claim. A class-balanced downsampling control preserves the order in 320/320 resamples (minimum per-class $n = 14$).

Era	Model	A/H/C	PC1 (%) RW / Int.	$\lvert\cos\rvert$ RW / Int.
Pre	Qwen	32/108/186	21.9 / 20.4	0.918 / 0.907
	Mistral	37/122/179	13.4 / 13.1	0.927 / 0.917
	Llama	19/120/177	51.6 / 52.0	0.990 / 0.990
	Gemma	62/107/179	20.2 / 20.0	0.985 / 0.982
Post	Qwen	168/92/14	39.0 / 37.7	0.931 / 0.901
	Mistral	163/78/21	22.3 / 21.4	0.979 / 0.981
	Llama	181/80/23	58.5 / 58.3	0.987 / 0.986
	Gemma	138/93/21	23.8 / 23.7	0.987 / 0.988

I Intervention Definitions

This appendix gives the exact edit definitions. Both edits start from the same class-mean difference: average abstention activations minus average hallucination activations. Dense applies layer-specific normalized versions of that difference in residual-write space. The sparse comparator applies the corresponding intermediate-space difference only on a fixed mask selected with the CETT contribution score.

For Dense Window Steering, each selected layer

ℓ uses

$$d_\ell = \mu_\ell^{A,\mathrm{dp}} - \mu_\ell^{H,\mathrm{dp}}, \qquad v_\ell = d_\ell / \|d_\ell\|_2.$$

At inference time, $o_\ell^{\mathrm{dp}} \leftarrow o_\ell^{\mathrm{dp}} + \alpha v_\ell$ for every $\ell \in L$, where o_ℓ^{dp} is the current down-projection output, L is the selected start-to-end layer range and α is edit strength. The vector is layer-specific: each selected layer has its own normalized A-H direction, and layers outside L are not edited. The layer range L and strength α are chosen during validation and then frozen for final evaluation.

For the sparse comparator, the fixed mask m_ℓ selects the intermediate neurons to edit. The mask is selected before validation. At the last prompt token, we compute a nonnegative CETT contribution-score vector, concatenate it across layers, standardize it, and train an L1 logistic probe to separate abstention from hallucination. Probe coordinates with coefficient magnitude above 10^{-6} form the fixed mask. For intermediate activation z_ℓ and down-projection matrix W_ℓ^{down}, the per-neuron selection score is

$$s_{\ell,j} = \frac{|z_{\ell,j}| \, \|W_{\ell,:,j}^{\mathrm{down}}\|_2}{\|W_\ell^{\mathrm{down}} z_\ell\|_2}.$$

The implementation clamps the denominator below at 10^{-8}. This score is the probe feature; the steering update itself uses the raw intermediate A-H difference on the selected coordinates, not the score or probe-coefficient magnitude. The intervention is

$$u_\ell = \mu_\ell^{A,\mathrm{int}} - \mu_\ell^{H,\mathrm{int}}, \qquad z_\ell \leftarrow z_\ell + \alpha(m_\ell \odot u_\ell),$$

where z_ℓ is the current intermediate activation and $\odot$ is elementwise multiplication. The update is zero outside the mask, so those activation coordinates remain unchanged.

J Intervention Alignment Checks

Dense Window uses validation-selected start-to-end layer ranges rather than all layers. Within the selected range, the implemented edit concatenates one unit-normalized A-H direction per layer, with equal layer weight before the scalar strength α is applied. Table 22 checks whether this actual range-limited edit remains aligned with the PCA direction: it reports the full-space A-H/PC1 cosine, the selected-range cosine, the per-layer cosine range, and the share of PC1 mass inside the selected range.

Table 22: **Dense Window alignment with PC1.** Global A-H $|\cos|$ is the Table 2 diagnostic computed in the full concatenated residual-write space. Dense-window $|\cos|$ compares PC1 with the actual equal-layer concatenation of per-layer unit edit vectors applied by Dense Window Steering; per-layer $|\cos|$ reports the local absolute alignment inside each edited layer. PC1 mass is the percentage of the global unit PC1's squared norm inside the selected layer range.

Metric	Qwen	Mistral	Llama	Gemma
Layer range	16–21	14–19	22–27	23–30
Global $\lvert\cos\rvert$	0.98	0.99	0.99	0.93
Window $\lvert\cos\rvert$	0.83	0.94	0.99	0.86
Layer $\lvert\cos\rvert$	0.89–0.97	0.98–0.99	0.99–0.99	0.95–0.96
PC1 mass (%)	10.1	12.9	15.5	34.2

The sparse comparator uses a sparse intermediate-neuron mask rather than a start-to-end layer range. Table 23 reports whether the actual masked edit remains aligned with the intermediate-space PCA direction, and how much of that PCA direction lies inside the selected mask.

Table 23: **Sparse-mask alignment with intermediate PC1.** The fixed intermediate-neuron mask is selected with an L1 probe trained on standardized CETT contribution-score features. Global intermediate A–H $|\cos|$ is computed in the full intermediate space. Masked-edit $|\cos|$ compares PC1 restricted to the mask with the masked raw A–H edit used for steering. Sparse-subspace $|\cos|$ refits PCA inside the selected mask. PC1 mass reports how much of full intermediate PC1 lies in the mask.

Metric	Qwen	Mistral	Llama	Gemma
Neurons	865	695	468	312
Mask layers	28	32	32	26
Global $\lvert\cos\rvert$	0.97	0.99	0.99	0.94
Masked $\lvert\cos\rvert$	0.99	1.00	1.00	0.98
Subspace $\lvert\cos\rvert$	0.98	1.00	1.00	1.00
PC1 mass (%)	0.545	0.729	0.153	0.223

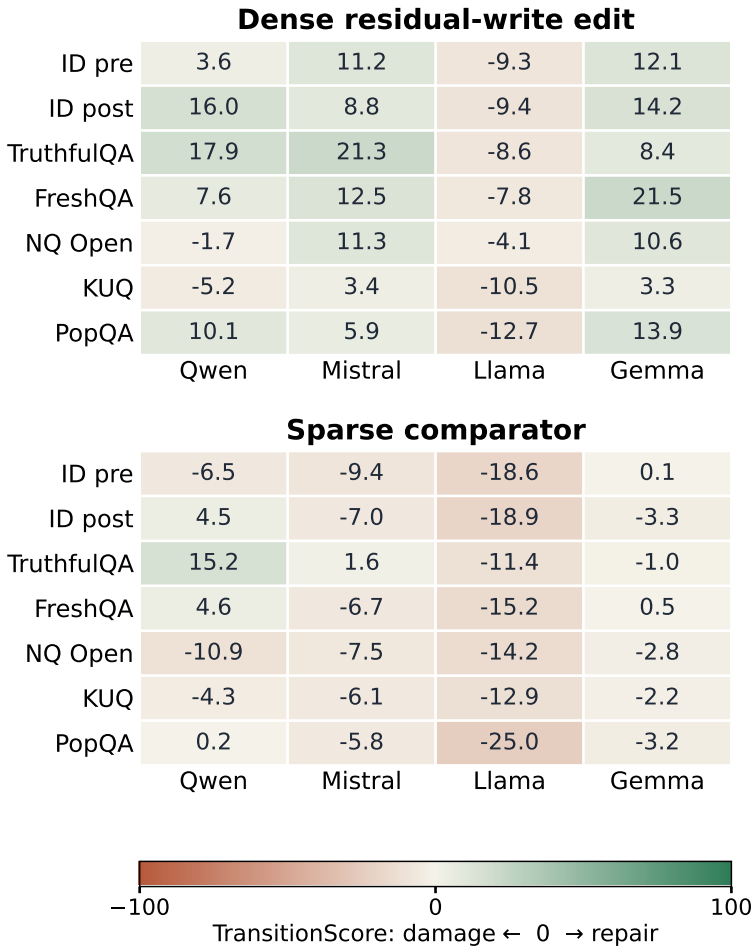


Figure 7: **Cross-model and cross-dataset TransitionScore overview.** Point estimates for validation-frozen Dense (top) and sparse comparator (bottom) edits. Positive values mean credited transitions outweigh penalties under this class-balanced utility; negative values mean the reverse. This secondary heatmap omits uncertainty. Table 3 reports raw Dense transitions, and Table 13 reports the corresponding sparse-comparator transitions. Table 10 gives Dense intervals across datasets; Table 12 gives exact sparse-comparator intervals and denominators for held-out ID-pre and ID-post.