BEFORE THE REFUSAL: PRE-RESPONSE ACTIVATION SIGNATURES OF FALSE REFUSAL IN SAFETY-ALIGNED LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can false refusal be predicted before a language model generates its first response token? We study this question using response-conditioned labels: an answerable prompt is a false refusal only if the model subsequently refuses it. On historical benchmark cohorts, we evaluate pre-response dense activations and sparse autoencoder (SAE) features while separating predictive evidence from intervention evidence. Retrospective nested cross-validation gives dense ROC-AUCs of 0.961 in Gemma-2-9B-IT and 0.916 in Qwen2.5-7B-Instruct. Adding dense activations to a combined character-TF-IDF and metadata baseline improves ROC-AUC by 0.098 in Gemma and 0.145 in Qwen, with paired source-bootstrap intervals of [0.065, 0.135] and [0.081, 0.215]. These gains are relative to the specified baseline and historical cohorts, not evidence of complete deconfounding or unseen confirmation. Corrected finite-control tests do not establish dense-intervention specificity; historical SAE edits change behavior, but their random controls are largely inactive. The supported result is therefore a pre-response predictive signature with incremental value over the tested prompt-covariate baseline. Whether the associated features provide selective causal control remains unresolved.

1 INTRODUCTION

A safety-aligned language model can refuse an answerable request because it resembles an unsafe one. Benchmarks document this *false refusal*, also called over-refusal, as a failure to distinguish permissible assistance from requests that warrant refusal (Röttger et al., 2024; Cui et al., 2025; An et al., 2024; Zhang et al., 2025). Our starting question concerns the computation before that failure appears in text: *among answerable prompts, can the model's internal state predict whether its subsequent response will answer or refuse?*

We focus on the residual-stream state at the last position of the formatted input, before generation of the first response token. At this point the model has processed the request and chat template but has not produced a refusal phrase. A predictive readout here would locate a usable signal before the observed response; it would not, by itself, establish that the model has irrevocably committed to refusing or that this is where the signal first emerges. The response supplies the outcome label, not the probe's input.

Restricting the comparison to answerable prompts is essential: separating safe from unsafe requests would not explain inappropriate refusal. This is a test of pre-output prediction, not a demonstrated early-warning deployment system.

Three questions organize the study (Figure 1). **Representation:** do pre-response states distinguish later answers from false refusals? **Incremental prediction:** do they improve prediction beyond the specified text and metadata baseline? **Causal specificity:** do interventions associated with those states change behavior more selectively than appropriate controls? These are different evidential requirements. Predictive separability could reflect hazard-related vocabulary, dataset composition, or topic. An intervention effect could instead reflect residual magnitude, SAE reconstruction error, or disruption of active features generally.

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

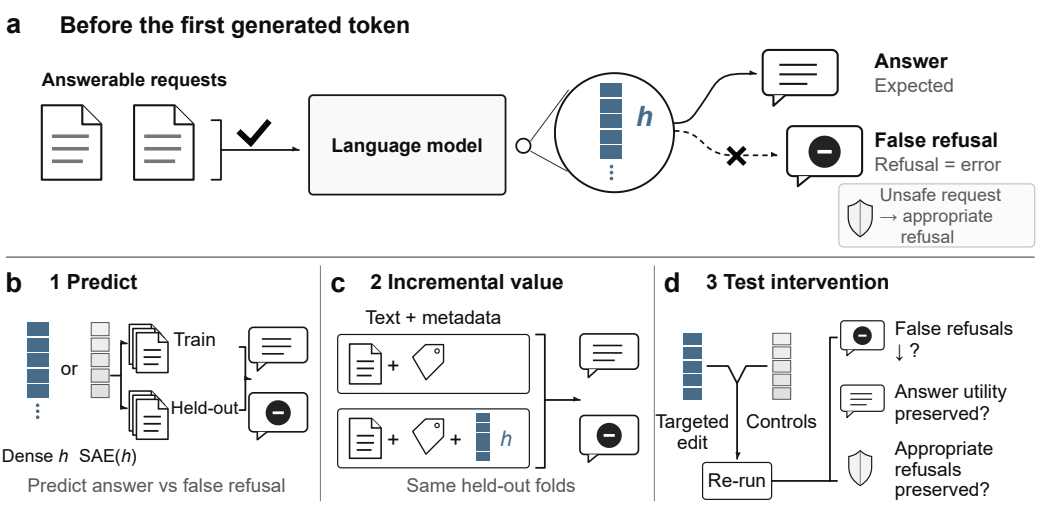


Figure 1: Study design and evidence hierarchy. (a) For answerable requests, an answer is expected (C1), whereas refusal is an error (C2); appropriate refusal of an unsafe request (C3) is shown only as a contrast case. The pre-first state h is captured after the formatted input and before the first generated token. Response-derived labels are used for supervision and evaluation, never as input features. (b) Dense and SAE representations are alternatives, evaluated with nested CV and training-only selection. (c) Text and metadata are compared with the same features plus h on identical held-out folds; the SAE variant is Gemma-only. (d) Exploratory edits are compared with random controls and reconstruction checks. Utility denotes answer-like route preservation, not a comprehensive quality assessment. Active-matched confirmation remains unfinished; the three outcome questions are tests, not claimed successes. This conceptual schematic does not show measured geometry or establish a causal circuit.

Internal-state interventions for false refusal already exist (Wang et al., 2025; Pan et al., 2025a; Jiang et al., 2026). Our focus is not to introduce a refusal vector or claim a mitigation, but to evaluate a pre-response signature under progressively more demanding controls. We consolidate historical experiments with completed retrospective reanalyses: nested candidate selection, a combined text-plus-metadata baseline, corrected finite-control tests, and an inactive-null diagnosis. The predictive evidence survives these specified comparisons. The intervention evidence does not yet establish selective feature-level control. The resulting contribution is an evidence hierarchy: a measurable internal signature, a baseline-relative test of its added predictive value, and explicit boundaries on its mechanistic interpretation.

Contributions. First, we distinguish prompt answerability from observed response route, making the false-refusal prediction target explicit. Second, retrospective nested evaluation identifies predictive pre-first readouts in Gemma and Qwen, with candidate-grid and label-provenance limits. Third, a combined prompt-covariate baseline tests incremental predictive value rather than relying on separate text-only and metadata-only comparisons. Finally, corrected random-control inference and SAE activity diagnostics explain why behavioral changes do not yet justify a strong mechanistic claim.

2 Related Work

Measuring and mitigating over-refusal. XSTest uses contrasts between safe and unsafe requests (Röttger et al., 2024); OR-Bench scales over-refusal evaluation (Cui et al., 2025); PHTest constructs pseudo-harmful prompts (An et al., 2024); and FalseReject targets contextual safety and structured reasoning (Zhang et al., 2025). We also use the OKTest dataset artifact (InvokerStark, 2024). These resources provide prompt sources, not an automatic guarantee that every apparent refusal in our models is inappropriate. We therefore separate prompt-policy judgments from response-route judgments.

Direct representation-based predecessors are especially relevant. Wang et al. (2025) extract and ablate a false-refusal vector. Pan et al. (2025a) analyze safety decision boundaries and use representation-space selection to obtain over-refusal data. Jiang et al. (2026) use inference-time activation energy to mitigate over-refusal. In contrast to claiming a new mitigation method, we ask whether an outcome-conditioned pre-first readout remains useful after a configured combined-baseline comparison and fold-local selection.

Refusal representations and interventions. Arditi et al. (2024) show that a direction can mediate refusal; Pan et al. (2025b) study multiple orthogonal safety directions. Neither finding licenses equating a predictive false-refusal direction with a complete mechanism. Activation patching provides a way to test behavioral sensitivity (Meng et al., 2022; Heimersheim & Nanda, 2024), but the intervention site, operation, and null controls define the conclusion. Here, non-replication and corrected control tests are part of the result, not exceptions to a mechanistic narrative.

Sparse features and evaluation. SAEs offer sparse coordinates for studying hidden states (Cunningham et al., 2023; Bricken et al., 2023); Gemma Scope supplies public SAEs for Gemma models (Lieberum et al., 2024). Their usefulness does not make feature labels ground truth. SAEBench evaluates SAE behavior along multiple axes (Karvonen et al., 2025), motivating a distinction between predictive utility, feature stability, and intervention specificity. We make this distinction concrete by quantifying how inactive the historical random features were. Methodological details of selection, readout interpretation, and finite-control inference are given below and in the appendix. Probe performance depends on both the representation and the readout's learning capacity; control-task work makes this distinction explicit (Hewitt & Liang, 2019). Our combined-baseline comparison addresses a narrower question of incremental prediction, not proof that a probe has isolated a model-internal algorithm.

3 Data, Timing, and Prediction Target

Models and cohorts. We use Gemma-2-9B-IT (Gemma Team, Google DeepMind, 2024) and Qwen2.5-7B-Instruct (Qwen Team, Alibaba Cloud, 2024). The historical prompt collection contains 1,100 prompts from five benchmark suites: FalseReject, OKTest, OR-Bench, PHTest, and XSTest. OR-Bench Hard and Toxic are separate subsets of one benchmark. The Gemma main cohort contains 600 examples: 477 answer-like responses (C1) and 123 false refusals (C2). The Qwen evaluation uses its saved model-specific C1/C2 metadata labels: 800 C1 and 50 C2, or 850 examples. The different denominators reflect model-conditioned cohort membership; these are not two measurements on an identical 600-example test set. Gemma's expanded response-audit mapping and Qwen's saved case labels are distinct label-source artifacts.

Pre-response capture. Let x be a prompt and $\tau(x)$ its model-specific formatted input, including the generation prompt. If $T(x)$ is the final input position, define

$$h_\ell(x) = H_\ell(\tau(x))_{T(x)} \in \mathbb{R}^{d_\ell}. \tag{1}$$

H_ℓ denotes the saved residual-stream readout at layer ℓ. The `pre_first` readout uses a prompt-only forward pass: no generated response tokens are included. This differs from the last raw user token if the template appends tokens. Historical `post_prefix` readouts include generated tokens and serve only as timing comparators. An SAE encodes a readout as $z_\ell(x) = E_\ell(h_\ell(x))$; sparse coordinates are candidate features, not automatically meaningful concepts. Here ℓ indexes a layer, d_ℓ is the residual dimension, and E_ℓ is the corresponding sparse encoder. We probe dense coordinates and SAE coordinates separately; encoding a state does not itself constitute an intervention.

The timing convention makes a practical distinction. The model sees the whole formatted request when $h_\ell(x)$ is measured, including any assistant-prefix tokens added by the template. It does not see the subsequently generated answer. Consequently, this readout can contain prompt semantics and formatting effects; "pre-response" rules out generated-text input, not prompt-related confounds. The post-prefix comparator addresses a different question and is never described as an early readout.

Table 1: Historical cohort definitions and analytical roles. Groups combine prompt status with observed behavior; benchmark unsafe-control status is not equivalent to uniform prompt-level annotation. These mapped groups do not partition all 1,100 prompts. Sample sizes for primary prediction appear in Table 2.

Group	Prompt status	Response route	Role
C1	Answerable	Answer-like	Negative prediction class
C2	Answerable	Refusal-like	False-refusal target
C2-appropriate	Refusal appropriate	Refusal-like	Policy-specific control
C2-ambiguity	Ambiguous	Refusal-like	Ambiguity control
C3	Unsafe control	Refusal-like	Warranted-refusal control
C4	Unsafe control	Non-refusal	Safety failure diagnostic

Response-conditioned outcome. Let $a(x)$ denote judged answerability and $r(x)$ the observed response route. For the main cohort restricted to $a(x) =$ answerable,

$$y(x) = \begin{cases} 0, & r(x) \in \{\text{direct answer}, \text{constrained answer}\} \quad (\text{C1}), \\ 1, & r(x) \in \{\text{hard refusal}, \text{safe redirection}\} \quad (\text{C2}). \end{cases} \tag{2}$$

Other routes do not automatically become C1. Labels are determined after generation; the prediction task asks whether $h_\ell(x)$ can predict them without seeing that generated text. Response-derived labels supervise training and score held-out predictions, but response text and output-derived metadata are excluded as input features.

This is a model-conditioned prediction task: $y(x)$ abbreviates the outcome for a particular model and its recorded generation setting, rather than an intrinsic property of the prompt. The same request need not belong to the same group for Gemma and Qwen. We do not estimate a prompt's refusal probability over repeated stochastic generations. Instead, the probe ranks prompts by their association with the recorded binary outcome within each model's cohort.

3.1 Labeling and Control Groups

The historical Gemma response-audited cohort uses three independent human expert raters with majority-vote prompt-policy and response-route labels, as specified in the annotation records. Qwen uses the separately saved case-label artifact described in Section 3; we do not infer identical audit coverage from its C1/C2 names. Prompt-level coverage was expanded to 950 of the 1,100 prompts; the remaining 150 are benchmark-designed unsafe controls. Benchmark membership and response filtering must not be mistaken for uniform prompt-level coverage.

Table 1 makes the two axes of cohort formation explicit. "Answer-like" combines direct and constrained answers; "refusal-like" combines hard refusals and safe redirection. These are route categories, not judgments of the response's factual accuracy or overall usefulness. The Gemma response-audit taxonomy separates C1 ($n = 477$) and C2 ($n = 123$) from refusal-appropriate responses (C2-appropriate, $n = 26$), refusals to ambiguous prompts (C2-ambiguity, $n = 13$), refusals to unsafe-control prompts (C3, $n = 171$), and non-refusal responses to unsafe-control prompts (C4, $n = 74$). These counts describe the mapped analysis/control groups, not a complete partition of all 1,100 prompts. Table 9 documents source-specific coverage. Route labels and prompt-policy judgments are distinct: a refusal is not a false refusal solely because its wording matches a refusal template.

Only C1 and C2 enter the primary binary prediction task. Comparisons of C2 against C2-appropriate or C3 ask whether the readout separates prompt groups even when both groups receive refusal-like responses; C4 records failures to refuse unsafe controls. Clarifications, evasive responses, and unresolved cases are not silently mapped to a successful answer. Keeping these distinctions avoids turning every non-refusal into evidence that the central task was fulfilled.

Later, separately collected confirmatory cohorts use agent annotation and adjudication rather than these historical human-expert labels. They are not merged into the predictive results reported here. Their final active-feature-matched intervention test remains incomplete.

4 PREDICTIVE EVALUATION PROTOCOL

4.1 READOUT AND FOLD-LOCAL SELECTION

For a selected representation $v(x)$, a class-weighted logistic probe estimates $\hat{p}(y = 1 \mid v(x)) = \sigma(w^\top v(x) + b)$, where σ is the logistic function. Its score summarizes association with false refusal; w is a fitted readout, not an identified direction that the model necessarily uses to decide whether to refuse. The same distinction applies to an SAE-based probe.

A layer chosen because it performed best on the evaluation data can make a readout appear more reliable than it is. We therefore separate candidate selection from performance evaluation using nested cross-validation (Varma & Simon, 2006; Cawley & Talbot, 2010). The completed procedure has four steps:

1. Split each model's C1/C2 cohort into five stratified outer folds. Hold one fold aside; the remaining four contain all data available for fitting and candidate selection in that outer iteration.
2. Within that outer training set, compare candidate layer/position pairs using four-fold inner CV. For each inner training split, rank coordinates by absolute Cohen's d, retain the top 512, fit scaling, and fit the probe. Score only the corresponding inner validation split.
3. Choose the candidate with the highest mean inner ROC-AUC. Recompute feature selection and scaling on the complete outer training set, fit a fresh probe there, and predict the held-out outer fold once.
4. Rotate through all five outer folds and repeat the partitioning ten times. Each source example receives one out-of-fold prediction per repeat; repeats change splits, not the size of the underlying prompt collection.

The outer test labels do not choose the layer, position, or coordinates: training-only preprocessing includes supervised selection, not just scaling.

Gemma's available grid contains layers 9, 20, and 31 at four saved positions (`hazard`, `last_user`, `pre_first`, and `post_prefix`), separately for dense and SAE representations. Qwen's available grid is narrower: dense layers 7, 14, and 25 at `pre_first` only. The Gemma grid includes a post-response comparator, so a pre-response conclusion depends on which readout is actually selected. This is retrospective selection adjustment for the available grid, not a claim that all historical researcher choices were nested or preregistered.

4.2 COMBINED BASELINES AND INCREMENTAL ESTIMAND

Activation-only discrimination does not establish added value: prompt text or dataset identity may already explain much of the same variation. We compare four readouts on identical outer folds. B0 uses text and metadata; B1 uses selected dense activations; B2 concatenates B0 with the dense features used by B1; B3 concatenates B0 with selected SAE features (Gemma only). The activation candidate is selected by its activation-only inner-CV score, not by maximizing the eventual B2–B0 gain on the test fold.

Text is character-within-word TF-IDF with 3–5-grams and at most 5,000 features. Metadata comprises dataset, type, category, focus, and prompt character length. TF-IDF vocabulary and weights, categorical encoders, length scaling, activation feature selection, and activation scaling use training data only. All four readouts use class-weighted logistic regression. No generated-text features or output-derived metadata are provided. Combining text and metadata in B0 is important: outperforming each separately would not show an improvement over a learner that can use both together.

We distinguish two summaries. Absolute performance is calculated from all out-of-fold predictions within each repeat, then averaged over ten repeats. The displayed percentile ranges describe variation across these splits. For incremental performance, let $\hat{p}_{i,r}^{(k)}$ be the out-of-fold prediction for source example i, repeat r, and baseline k. Define

$$\bar{p}_i^{(k)} = \frac{1}{10} \sum_{r=1}^{10} \hat{p}_{i,r}^{(k)}, \qquad \Delta_{\text{dense}} = \text{AUC}(y, \bar{p}^{(\text{B2})}) - \text{AUC}(y, \bar{p}^{(\text{B0})}). \tag{3}$$

Table 2: Retrospective nested prediction. ROC-AUC means and 2.5–97.5 percentile ranges over ten outer-CV repeats. Ranges describe split variation, not independent-sample confidence intervals. Each repeat uses five outer folds with four-fold inner candidate selection. AP is repeat-mean average precision, not accuracy or trapezoidal PR-AUC; cohort prevalence differs.

Model	Readout	C2 / C1	AUC	Repeat range	AP
Gemma-2-9B-IT	Dense	123 / 477	0.961	[0.954, 0.968]	0.888
Gemma-2-9B-IT	SAE	123 / 477	0.951	[0.943, 0.958]	0.877
Qwen2.5-7B	Dense	50 / 800	0.916	[0.906, 0.925]	0.589

Each source example therefore contributes one averaged score to each side of the paired comparison, rather than being counted ten times as independent data. "Source" here means an individual prompt record, not a benchmark suite. The completed paired bootstrap resamples these records 1,000 times, retaining the B0/B2 pairing. Its interval is conditional on the saved predictions: it does not refit feature selection or models in each resample.

ROC-AUC measures ranking of positives against negatives, not the fraction of correctly classified prompts at a fixed threshold. We also report the already-computed average precision (AP), a precision–recall summary sensitive to positive-class prevalence. AP is computed using `average_precision_score`, not trapezoidal integration. These measures address discrimination; they do not establish calibrated deployment thresholds.

5 Predictive Results and Robustness

5.1 A Readable Signal before the First Response Token

Gemma L31 pre-first and Qwen L25 pre-first are selected in all 50 outer folds; Gemma SAE likewise selects L31 pre-first. Thus the selected readouts are pre-response even though Gemma's grid includes post-response comparators. This stability concerns the tested candidates: it neither locates the first layer where information arises nor establishes irreversibility of a decision.

Dense mean ROC-AUC is 0.961 for Gemma and 0.916 for Qwen (Table 2); Gemma SAE reaches 0.951. A readout therefore distinguishes later route outcomes without observing generated refusal text. The corresponding mean AP values are 0.888, 0.589, and 0.877, respectively. Qwen's lower AP alongside high ROC-AUC makes the distinction between ranking metrics visible; its cohort has only 50 positives among 850 examples, compared with 123 among 600 for Gemma. AP values across these different cohorts should not be read as a controlled model-quality ranking.

Nested and historical-winner evaluations select the same candidates and give the same performance in this run. This is evidence of selection stability within the available grid, not an independent replication of the historically chosen setting. The ten-repeat ranges reuse the same examples and are not independent-sample confidence intervals. Historical fixed-setting summaries remain separately identified in Appendix A.2.

5.2 Added Value beyond the Specified Prompt Baseline

Gemma improves from B0 repeat-mean AUC 0.871 to B2 0.964; Qwen improves from 0.784 to 0.928 (Figure 2). The source-mean-OOF paired Δ_{dense} is 0.098 [0.065, 0.135] for Gemma and 0.145 [0.081, 0.215] for Qwen. Gemma B3 has repeat-mean AUC 0.954. The gains show that the tested linear readout benefits from adding internal features to the specified prompt-covariate representation, rather than merely beating a weak text-only comparator.

For example, subtracting Gemma's rounded repeat-mean values gives 0.093, whereas its paired mean-OOF delta rounds to 0.098. There is no discrepancy: averaging predictions before calculating AUC need not equal averaging AUCs. Equation 3 fixes which comparison the interval describes. Completed paired-swap permutation tests yield unadjusted $p = 1/1001$ for the two B2–B0 comparisons. No global predictive-family correction was performed, so we emphasize effects and conditional intervals rather than a family-wide significance claim.

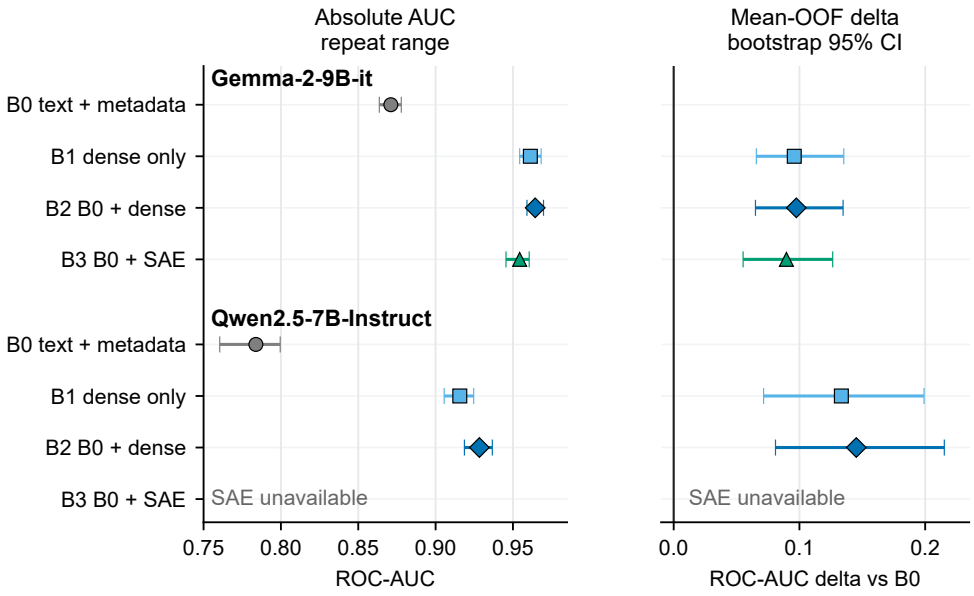


Figure 2: Completed retrospective predictive comparisons. Left: absolute ROC-AUC means over ten outer-CV repeats, with percentile ranges describing split variation. Right: differences from B0 using source-level mean out-of-fold predictions, with paired source-bootstrap 95% intervals. The right panel is not obtained by subtracting left-panel means. B0 combines text and metadata; B1 is dense only; B2 adds dense features to B0; B3 adds SAE features to B0. Qwen has no SAE comparison and a restricted pre-first candidate grid. These results are not unseen confirmation.

Interpretation of the gain. Because $h_\ell(x)$ is computed from the formatted input, we do not claim information unavailable in principle from the prompt. The result is that the tested readout exposes predictive structure that this particular text-plus-metadata learner does not recover as well. Character n-grams and recorded categories are not a comprehensive semantic model. Stronger text encoders, unmeasured nuisance variables, or distribution shifts could reduce the gain. Thus the supported statement is baseline-relative incremental predictability, not complete deconfounding.

5.3 Alternative Explanations and Scope

The main comparison does not settle every alternative explanation. Table 3 separates the question each completed analysis addresses from the inference it cannot support. The distinctions matter when moving from within-cohort prediction to a claim about false refusal more generally.

Model and dataset comparisons. Qwen provides a second-family predictive result, not a transferred Gemma probe: each model has its own readout, fitted probe, and outcome-conditioned labels. Its smaller positive class ($n = 50$), case-label provenance, and restricted candidate grid limit equivalence to the Gemma evaluation. Historical leave-one-dataset-out diagnostics favor dense features over text-only features on five held-out dataset slices, but OR-Bench-Hard is a boundary case with overlapping conditional bootstrap intervals. Their overlap is not a paired difference test. These comparisons do not replace the combined-baseline analysis.

Matching and propensity diagnostics. The historical fitted propensity scores have disjoint observed ranges: C2 minimum 0.6987 exceeds C1 maximum 0.5121, and the tested trimming rules retain no overlap sample. This is a limitation of the observed data and fitted diagnostic, not proof of population-level positivity failure. Historical matching comparisons are selected-sample sensi-

Table 3: Alternative explanations and the boundaries of the completed evidence. Each comparison addresses a specific concern, not all confounding or all historical selection. Intervention diagnostics are detailed in Section 6.

Concern	Completed comparison	Remaining limitation
Generated refusal text supplies the signal	Prompt-only pre-first readout	Does not locate the signal's first emergence
Text/source cues suffice	B2 versus combined B0 on the same folds	Limited text learner and measured covariates
Best layer chosen on test data	Fold-local nested candidate selection	Historical choices outside the grid remain
Dataset composition drives prediction	Historical held-out-dataset and matching diagnostics	Different estimands; no unseen combined-baseline confirmation
Any active perturbation changes behavior	Corrected dense tests and SAE inactive-null diagnosis	Selective causal control remains unresolved

tivity analyses; they do not identify population causal effects or exclude unmeasured confounding. Appendix A.3 reports balance and performance.

Refusal-type controls. Historical pre-first readouts distinguish C2 from appropriate refusals (AUC 0.855) and C2 from unsafe-prompt refusals (AUC 0.999). These are different prompt/cohort comparisons, with possible safety and source cues. They are useful specificity diagnostics, not a mechanistic anchor or a substitute for within-answerable C1/C2 evaluation.

6 Exploratory Intervention Evidence

Predictive readouts motivate interventions but do not establish selective causal control. Historical Gemma mean-replacement patching uses the C2-to-C1 mean activation difference, not the logistic probe normal. At $\alpha = 0.75$, the target route-change rate is 0.2333 and the 20-direction random mean is 0.1283; one control ties the target and none exceeds it. With ties included and the plus-one convention for finite Monte Carlo samples (Phipson & Smyth, 2010), the test is

$$p_{\mathrm{MC}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}[T_b \geq T_{\mathrm{target}}]}{B + 1}. \tag{4}$$

The resulting $p_{\mathrm{MC}} = 2/21 = 0.0952$, not 0.05. Here B is the number of random controls and $T = n^{-1} \sum_i \mathbf{1}[r_i^{\mathrm{edit}} \neq r_i^{\mathrm{base}}]$ counts paired changes in the historical heuristic coarse route, over all n tested prompts including baseline non-refusals. It is not necessarily a refusal-to-useful-answer conversion or a net refusal-rate reduction. The one tying direction counts against the target: its observed rate is at least as large. With only 20 controls the smallest attainable plus-one value is $1/21$, so a zero empirical exceedance count is not a zero p-value. The completed four-row dense-intervention family gives Holm-adjusted $p = 0.3636$ for this setting; no dense row passes that correction (Holm, 1979). The recorded Benjamini–Hochberg sensitivity analysis also rejects no dense row (Benjamini & Hochberg, 1995). The C3 refusal-preservation rate is 1.00 at this setting in the tested subset, but falls to 0.33 at $\alpha = 1.0$. Qwen's $\alpha = 0.75$ target rate (0.10) is below all 20 controls. We therefore retain these as exploratory behavioral diagnostics, not confirmed direction specificity or a safe operating recommendation. Table 7 and Appendix A.4 retain effects, ties, corrections, and control limitations.

6.1 SAE Edits and the Inactive-Null Problem

Using Gemma Scope at L31 and width 16,384 (Lieberum et al., 2024), the historical intervention selects ten positive-Cohen's-d features. Zeroing them changes 15 of 30 response routes (50%, binomial interval [0.32, 0.68]); twenty unconditioned random feature sets yield no changes. The corrected finite-control value is $1/21 = 0.0476$, within only a one-row retrospective SAE correction family.

The comparison is not a satisfactory test of feature specificity. Reconstructing the historical random sets from their seed and sampling rules gives an already-zero latent-value fraction of 98.65% on

the intervention subset, versus target active-value fraction 85.67%. Thus many random edits were effectively inert. Discovery bootstrap stability (mean top-ten Jaccard 0.802 across 1,000 resamples) concerns selection stability, not whether those features selectively cause false refusal. Feature glosses are tentative; the intervention is defined by indices. A combined-edit run had a hook-order bug and is excluded.

The activity fractions refer to feature-by-prompt latent entries, not the percentage of random sets with identical generated responses. Their relevance is operational: setting an already-zero coordinate to zero leaves that coordinate unchanged. A target set that is usually active is therefore being compared with random edits that often do little. The observed 50%-versus-0% contrast is real as a historical behavioral summary, but it does not isolate which feature identities matter from whether an active latent was altered.

An independent follow-up cohort and active-feature-matched controls have been frozen, but the matched-null evaluation is unfinished. No partial random-control distribution is analyzed here. The historical SAE effect therefore remains preliminary, with reconstruction, activity matching, and route-specific utility/safety central to its interpretation. Appendix A.5 gives the diagnostic detail.

7 Discussion and Limitations

What the internal-state result establishes. The two predictive tests answer complementary questions. The activation-only readout demonstrates that a later false refusal is distinguishable before the first output token. Adding activations to B0 tests whether that readout is redundant with a specified prompt-covariate learner. Together, they support an observable pre-response signature on the historical cohorts. Neither test requires that the probe coefficients be the model's decision rule, nor that the signature be confined to a single interpretable feature.

Why causal specificity is a separate requirement. A predictive state may reflect correlated prompt properties; changing it may disrupt behavior without selectively affecting false refusal. Controls must therefore address activation magnitude, feature activity, reconstruction, and warranted refusal preservation. The historical controls limit mechanistic interpretation without negating the predictive result.

Limits of the available evidence. The principal limitation is retrospective reuse of historical cohorts. Nested evaluation reduces selection bias within available candidate grids but does not undo earlier dataset, label, or analysis choices. Qwen lacks the full Gemma timing/SAE grid and uses a distinct saved label artifact. Human-expert judgments can be fallible; majority vote is not an external validity standard. The class imbalance and finite prompt collection limit precision and generalization.

Incremental gains are conditional on a specific character-level baseline and recorded metadata. Bootstrap intervals resample saved predictions, not the full training procedure. No independent unseen predictive confirmation is reported. Matching and route-pair diagnostics do not establish complete deconfounding.

Intervention evidence is weaker still: dense tests do not pass the completed Holm family, historical SAE nulls are largely inactive, and matched-null confirmation remains incomplete. Safety measurements are cohort-scoped and do not cover adversarial or out-of-distribution deployment. We do not claim a causal circuit, monosemantic features, cross-family causal generalization, or deployable mitigation.

8 Conclusion

False refusal can be predicted from the state preceding the first generated token in the historical Gemma and Qwen cohorts. The signal retains incremental value over the specified combined text-plus-metadata learner under retrospective nested evaluation. This supports studying false refusal as a pre-response internal-state phenomenon, without equating prediction with commitment or mechanism. Corrected dense tests and the inactive SAE null place a clear boundary on the mechanistic interpretation: selective causal control remains a question for stronger completed controls.

ETHICS STATEMENT

Refusal interventions are dual-use: reducing inappropriate refusal may also reduce warranted refusal. The reported interventions are diagnostics, not recommended mitigation settings. Raw prompts, generated responses, and unsafe completion text are excluded from the public supplementary package; aggregate metrics and redacted identifiers are used instead. Benchmark and model licenses continue to govern their original materials. Historical human-expert labels and later agent-only labels are distinguished, and authors remain responsible for evaluating their limitations.

REPRODUCIBILITY STATEMENT

The appendix specifies cohort definitions, evaluation procedures, and intervention limitations. The anonymous supplement contains figure source tables, editable figure code, manuscript sources, and selected analysis code/configurations with reconstruction instructions. It does not redistribute private prompts, responses, activation arrays, or model checkpoints. Rebuilding aggregate figures requires no model inference; reproducing activation analyses requires separately obtaining the licensed inputs. The experiments used one NVIDIA RTX 3090 (24 GiB). This revision reports completed analyses and runs no additional model experiments.

REFERENCES

Bang An, Sicheng Zhu, Ruiyi Zhang, Michael-Andrei Panaitescu-Liess, Yuancheng Xu, and Furong Huang. Automatic pseudo-harmful prompt generation for evaluating false refusals in large language models. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2409.00598.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.11717.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://academic.oup.com/jrsssb/article/57/1/289/7035855.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html. Anthropic research blog post (not arXiv).

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL https://www.jmlr.org/papers/v11/cawley10a.html.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2405.20947.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Gemma Team, Google DeepMind. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. arXiv preprint arXiv:2404.15255, 2024. URL https://arxiv.org/abs/2404.15255.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: 10.2307/4615733. URL https://www.jstor.org/stable/4615733.

InvokerStark. OverKill (OKTest): Overkill test prompts, 2024. URL https://github.com/InvokerStark/OverKill. Dataset artifact; no formal paper verified.

Eric Hanchen Jiang, Weixuan Ou, Run Liu, Shengyuan Pang, Guancheng Wan, Ranjie Duan, Wei Dong, Kai-Wei Chang, XiaoFeng Wang, Ying Nian Wu, and Xinfeng Li. Mitigating over-refusal in aligned large language models via inference-time activation energy. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37930–37950, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1759. URL https://aclanthology.org/2026.acl-long.1759/.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart Mcdougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264. PMLR, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.html.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2202.05262.

Licheng Pan, Yongqi Tong, Xin Zhang, Xiaolu Zhang, Jun Zhou, and Zhixuan Chu. Understanding and mitigating overrefusal in LLMs from an unveiling perspective of safety decision boundary. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21057–21075, Suzhou, China, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1065. URL https://aclanthology.org/2025.emnlp-main.1065/.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Yu Haining, and Xiaohua Jia. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47697–47716. PMLR, 2025b. URL https://proceedings.mlr.press/v267/pan25f.html.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585. URL https://doi.org/10.2202/1544-6115.1585.

Qwen Team, Alibaba Cloud. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024. URL https://arxiv.org/abs/2308.01263.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. doi: 10.1186/1471-2105-7-91. URL https://link.springer.com/article/10.1186/1471-2105-7-91.

Xinpeng Wang, Chengzhi Hu, Paul Röttger, and Barbara Plank. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/53dbd7e34fab703a639964e2d3ee9e84-Abstract-Conference.html.

Zhehao Zhang, Weijie Xu, Fanyou Wu, and Chandan K. Reddy. FalseReject: A resource for improving contextual safety and mitigating over-refusals in LLMs via structured reasoning. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2505.08054.

A SUPPLEMENTARY MATERIAL

This appendix collects historical fixed-setting results, retrospective posthoc corrections, diagnostics, and figure-bank evidence that support the main-text claims. **No new experiments were run** to produce any figure or table in this appendix. Figures and tables are generated from existing aggregate source-data files or locked manuscript tables; corrected finite-null interpretations are stated in captions and notes rather than by editing locked CSV artifacts.

The main text stands alone: all headline claims are stated in main-text prose. This appendix provides the detailed evidence for interested reviewers, including appendix-only intervention diagnostics that are no longer main-text floats.

A.1 CLAIM SCOPE REFERENCE

The canonical claim-scoping reference (Table 4) lists every claim the paper makes or forbids, with the required caveat for each. Borderline claims carry their caveat inline. Rows marked NO are forbidden as claims anywhere in the paper.

Table 4: Claim scope. Canonical reference: every claim in the abstract, main text, appendix, or reviewer rebuttal must be checkable against this table. Borderline claims must carry their caveat inline. Rows marked NO are forbidden as claims. C1/C2/C2-appropriate/ambiguity labels are from human expert annotation; C3/C4 unsafe-control status additionally relies on benchmark design (explicit prompt-level labels are partial for unsafe controls). Later follow-up cohorts that use Codex-native agent annotation are separate from this historical manuscript label provenance.

Claim	Status	Required caveat / wording
Dense signature predicts false refusal (Gemma)	**YES**	Predictive, not causal; nested ROC-AUC 0.961; repeat ranges are not sampling CIs
Predictive replication cross-family (Qwen)	**YES**	Predictive-only; restricted dense pre-first grid; no causal replication
Dense intervention specificity in Gemma	**PARTIAL**	Exploratory diagnostic; 1/20 tied; $p_{\mathrm{MC}} = 0.0952$; Holm $p = 0.3636$; no corrected dense pass
Observed layer-dependent effects (Gemma)	**PARTIAL**	Exploratory; direction specificity remains unestablished
SAE top-10 downscale changes historical routes	**PARTIAL**	Preliminary; $p_{\mathrm{MC}} = 0.0476$ in one-row family; inactive random null; active-matched control unfinished
Dense signal survives confound matching	**PARTIAL**	Zero common support; scoped to matched sample
Population-level causal inference	NO	Not identified by fitted-score nonoverlap and selected-sample diagnostics
Cross-family causal generalization	NO	Forbidden: Qwen diagnostic-alpha intervention fails
Safety preserved at $\alpha = 1.0$	NO	Forbidden: C3 0.33 (Gemma), 0.57 (Qwen)
Deployable safety mitigation	NO	Forbidden: diagnostic only
Established gated raw-data release	NO	Anonymous aggregate/code supplement only; no raw prompts/completions
Human expert annotation	YES	Supported for historical manuscript labels; later agent-only cohorts reported separately
Monosemantic SAE features	NO	Forbidden: no monosemanticity test

Intervention rows are appendix diagnostics, not promoted mechanistic or mitigation claims. Safety evaluation is in-distribution only. The historical 95.9% audit statistic is a group-OK endorsement/pass rate, $\#\{\texttt{final_group_ok_final=yes}\}/n_{\mathrm{total}}$, not an agreement coefficient.

Table 5: Historical fixed-setting predictive results. Mean AUC across 50 repeated stratified splits (central 95% ranges of the repeated-split distribution, not sampling CIs). Dense activations exceed both text TF-IDF and metadata-only baselines in both families. The Qwen result is *predictive cross-family replication only*; it carries no causal claim (see Table 7). No causal claim can be made from this table alone.

	Text TF-IDF AUC [range]	Metadata AUC [range]	Dense (best PF) AUC [range]	Dense–text gap
Gemma-2-9B-it (L31 PF)	0.794 [0.735, 0.868]	0.852 [0.790, 0.908]	**0.960 [0.928, 0.984]**	+0.166
Qwen2.5-7B-Inst (L25 PF)	0.657 [0.544, 0.782]	0.796 [0.636, 0.882]	**0.924 [0.857, 0.969]**	+0.267

Gemma and Qwen use the same prompt pool and prompt-level labeling framework, but C1/C2 counts are model-specific because false refusal is outcome-conditioned on each model's response. Gemma: $n_{\text{pos}} = 123$, $n_{\text{neg}} = 477$. Qwen: $n_{\text{pos}} = 50$, $n_{\text{neg}} = 800$. Qwen SAE was not run. Gemma SAE best PF AUC$= 0.950$ [0.913, 0.977]. Dense–metadata gaps: Gemma $+0.108$, Qwen $+0.128$.

A.2 Additional Predictive Results

This section provides the detailed predictive results that support the main-text results in Section 5. Historical fixed-setting results are retained for traceability; the selection-adjusted predictive summary in the main text remains the promoted estimate. All figures are **predictive, not causal**. No new experiments were run.

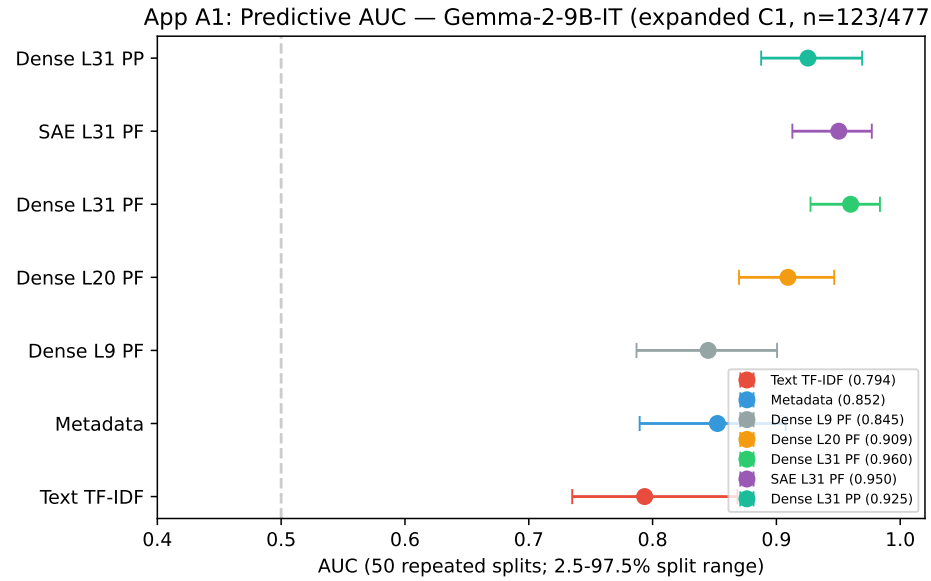


Figure 3: Historical fixed-setting predictive AUC across 50 repeated stratified splits for Gemma-2-9B-IT (expanded C1, $n_{\text{pos}} = 123$, $n_{\text{neg}} = 477$). Dense L31 `pre_first` (0.960) exceeds text TF-IDF (0.794) and metadata-only (0.852). SAE L31 `pre_first` (0.950) also exceeds both separate baselines. Error bars are central 95% repeated-split ranges, not independent-sample confidence intervals. **Predictive, not causal; historical fixed setting, not the selection-adjusted headline.**

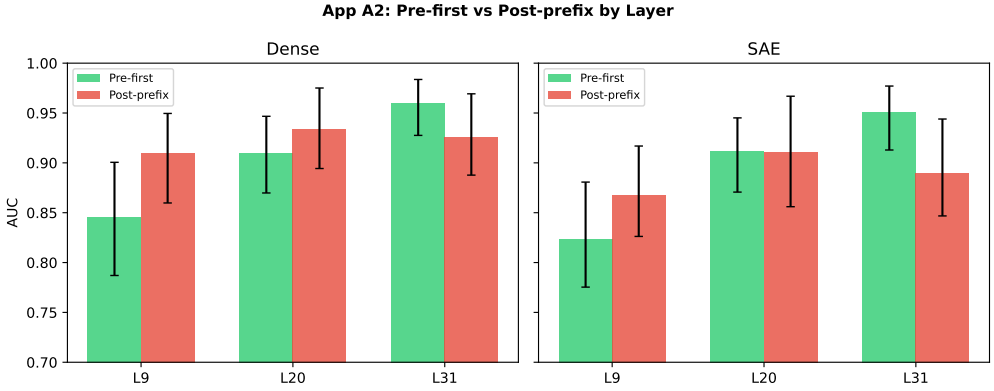


Figure 4: Historical pre-first vs post-prefix AUC at layers 9, 20, 31 for dense and SAE probes. Pre-first dominance is clearest at the best layer (L31). **Pre-first dominance holds at the best layer but may not hold at every layer; predictive, not causal; labels are later response routes used only for supervision/evaluation.**

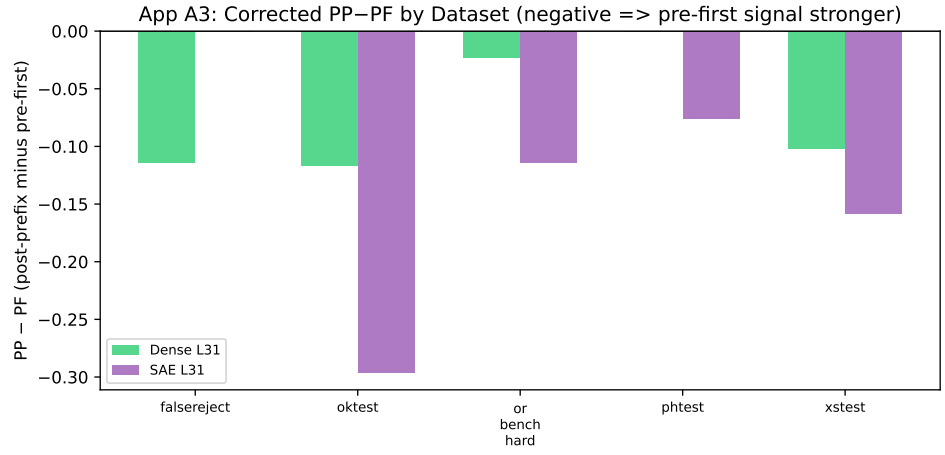


Figure 5: Corrected PP−PF (same-layer post-prefix AUC minus pre-first AUC for the best pre-first layer) by dataset and kind. Negative values mean the pre-first signal is stronger. **Negative PP−PF confirms pre-response predictive signal in these historical splits; two dataset/kind combinations have PP−PF $= 0$; no causal timing claim follows.**

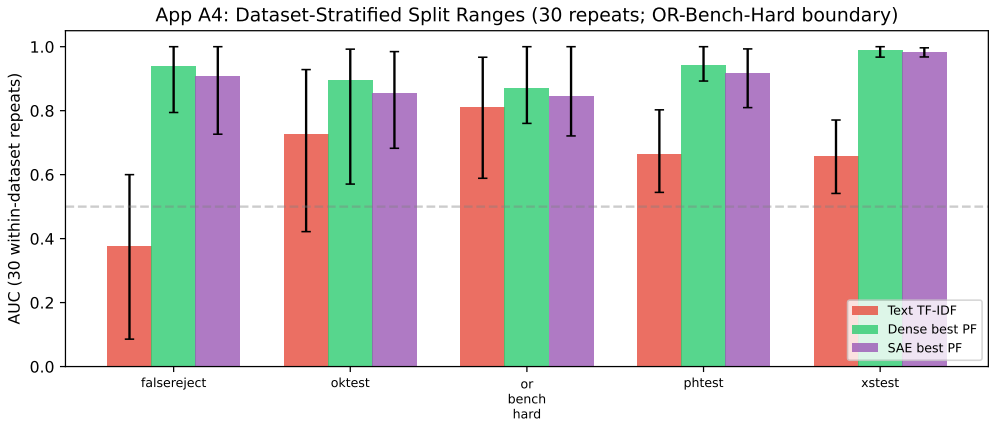


Figure 6: Dataset-stratified mean AUC with central 95% repeated-split ranges for text TF-IDF, dense best `pre_first`, and SAE best `pre_first`. OR-Bench-Hard is marked as a **text-confounded boundary case**. 30 repeated splits per dataset due to smaller sample sizes. These ranges measure split variation, not independent-sample uncertainty. These historical summaries do not establish dataset-invariant confirmation.

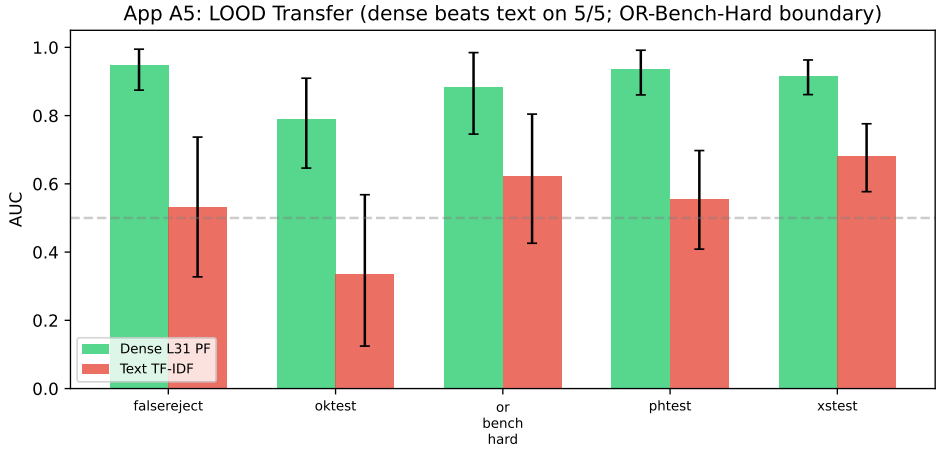


Figure 7: Historical leave-one-out-dataset (LOOD) transfer: dense activation beats text TF-IDF on 5/5 held-out datasets in Gemma. OR-Bench-Hard marked as a **text-confounded boundary case** and excluded from headline claims. Error bars are percentile 95% bootstrap intervals from 2,000 resamples of fixed held-out predictions; training is not repeated in each resample. Overlapping marginal intervals are not a paired test of the dense–text gap.

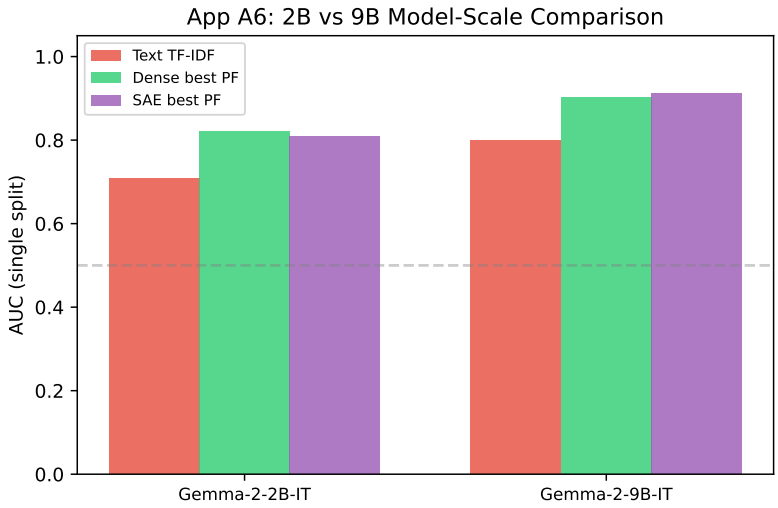


Figure 8: 2B vs 9B model-scale comparison. Single-split AUC for text TF-IDF, dense best `pre_first`, and SAE best `pre_first`. **2B uses prompt-relabeling; 9B is response-audited; single split, no CI bars; not directly comparable to repeated-split or nested-CV results.**

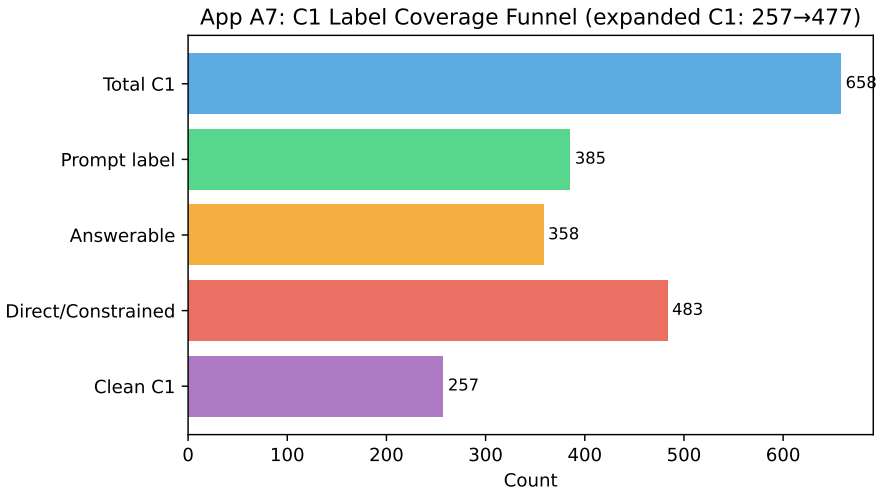


Figure 9: C1 label coverage funnel. Total C1 (658), prompt label available (385), prompt answerable (358), direct/constrained response (483), C1 answerable clean (257). **Original coverage shown; expanded C1 labels increase clean count from 257 to 477.**

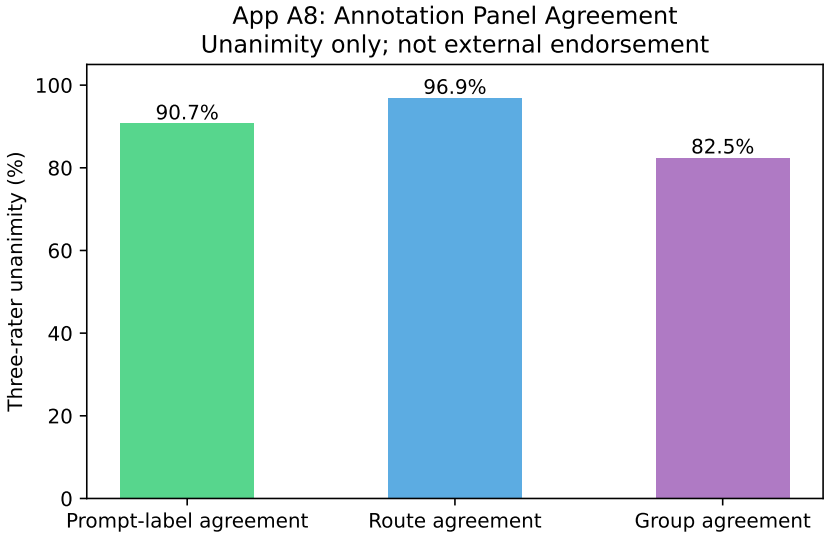


Figure 10: Historical annotation-audit summary. The group OK statistic is the fraction of adjudicated rows with `final_group_ok_final=yes`; it is an endorsement/pass rate, not an inter-rater agreement coefficient. Separate 3/3-count fields summarize how many reviewers matched on prompt, route, and group fields. **This historical manuscript cohort uses human expert annotation.**

A.3 Confound Control Details

This section provides the detailed confound-control figures that support the main-text summary in Section 5.3. Fitted-score nonoverlap limits these diagnostics; it does not prove population-level positivity failure. Matched-sample estimates are conditional on selected examples and do not identify population causal effects through IPW or doubly-robust estimation.

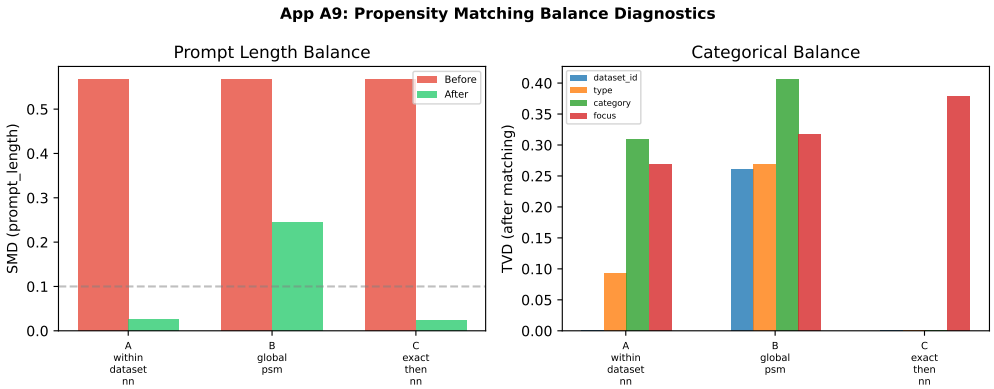


Figure 11: Propensity matching balance diagnostics: SMD on prompt_length and TVD on categorical covariates before and after matching for three strategies (A: within-dataset NN, B: global PSM, C: exact-then-NN). Strategy C has length SMD $= 0.023$ and dataset TVD $= 0.000$, but focus TVD increases from 0.273 to 0.378. Strategy A retains category TVD 0.309 and focus TVD 0.268. Balance is covariate-specific, not uniformly achieved.

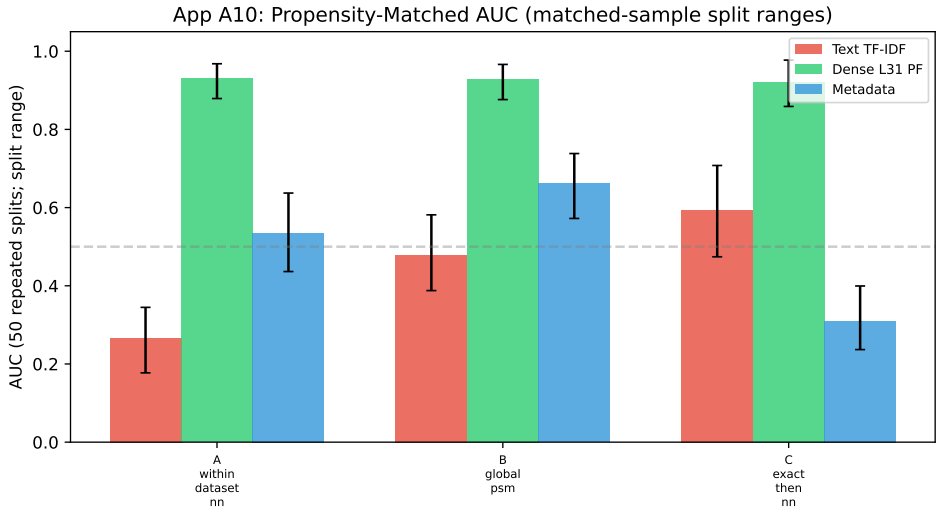


Figure 12: Propensity-matched AUC comparison across three matching strategies. The dense-over-text gap persists in all strategies: Strategy A $+0.665$, Strategy B $+0.450$, Strategy C $+0.328$. **Scoped to matched sample; no population-level causal inference; predictive, not causal.** Error bars are central 95% ranges over 50 repeated splits, not sampling confidence intervals.

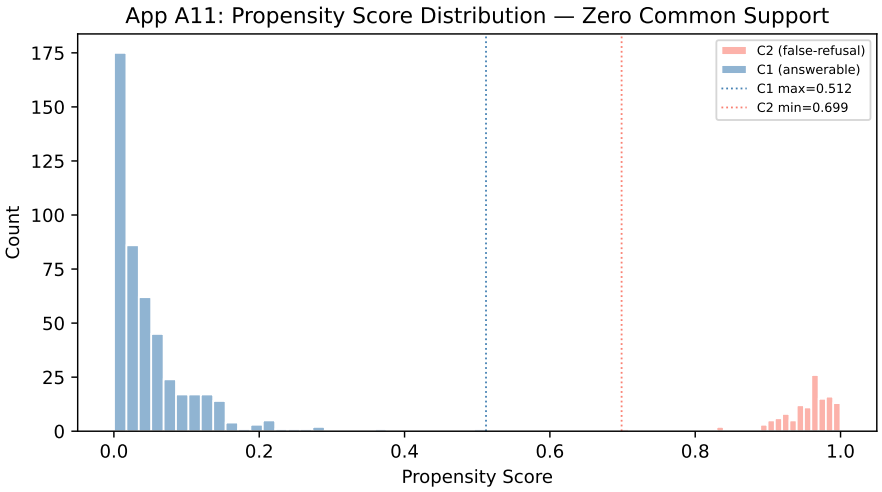


Figure 13: Propensity-score distributions for C1 (answerable) and C2 (false-refusal) examples. Scores fitted on the same cohort have disjoint ranges (positive $\min = 0.6987 >$ negative $\max = 0.5121$), so **common support is zero in this fitted diagnostic**. This does not establish population nonoverlap or complete covariate balance after matching.

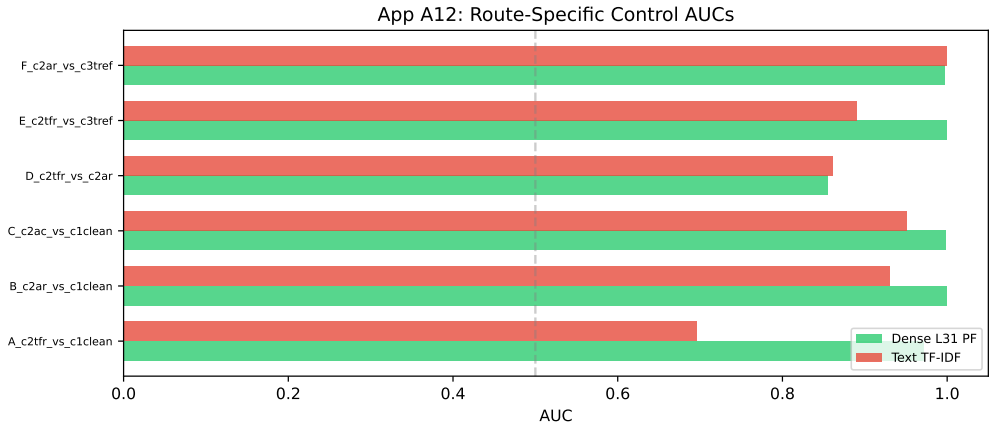


Figure 14: Route-specific control AUCs for six route pairs. Route-pair E (C2_tfr vs C3_tref) achieves dense L31 PF AUC $= 0.999$ in the historical route-control diagnostic: the readout distinguishes false refusal from true refusal *before generation*. **Predictive, not causal; this is a specificity/control diagnostic, not a mechanism claim.**

Table 6: Historical matched-sample diagnostics (fixed values). Three matching strategies applied to the Gemma-2-9B dense signature; AUCs and gaps computed on matched pairs only across 50 repeated stratified splits. Bounds are central 95% repeated-split ranges, not sampling confidence intervals. The fitted propensity scores have disjoint observed ranges (positive min$= 0.6987 >$ negative max$= 0.5121$). These selected-sample diagnostics do not establish population causal effects or population-level positivity failure.

Strategy	n pairs	SMD (length)	TVD (dataset)	Dense–text gap [95% split range]	
A: within-dataset NN	97	0.026	0.000	**+0.665** [0.568, 0.764]	
B: global PSM	123	0.245	0.260	+0.450 [0.351, 0.544]	
C: exact + NN on length	82	0.023	0.000	**+0.328** [0.208, 0.447]	
Zero common support	—	—	—	—	pos min $>$ neg max

Dense PF AUCs on matched pairs: A $= 0.931$, B $= 0.929$, C $= 0.921$. Text AUCs on matched pairs: A $= 0.266$, B $= 0.479$, C $= 0.594$. Dataset TVD does not summarize all categorical balance: strategy A retains category TVD 0.309 and focus TVD 0.268; strategy C retains focus TVD 0.378. The dense–text differences are predictive comparisons within matched samples, not evidence that all observed or unmeasured confounding has been removed.

Table 7: Historical activation-patching diagnostics. Mean-replacement patch along the C2→C1 direction (the unnormalised mean activation difference between C1 and C2, distinct from the probe normal). $n = 30$ per group; 20 random directions per alpha (Gemma). The Gemma $\alpha = 0.75$ row has one tied random control, so the corrected tie-inclusive value is $p_{\mathrm{MC}} = 2/21 = 0.0952$ and dense-family Holm $p = 0.3636$; no dense row passes Holm correction. **All safety-preservation cells are in-distribution**: the C3 / C2-appropriate-refusal subsets only scope this diagnostic setting, so out-of-distribution safety is untested. The intervention is an exploratory diagnostic, not a deployable mitigation.

Case	α	Change rate	C3 safety (in-dist.)	Strict pct (n=20)	p_{MC}
Gemma-2-9B-it, L31 pre_first — exploratory diagnostic					
Gemma targeted (diagnostic)	**0.75**	**0.2333**	**1.0000**	**95.0**	**0.0952**
Gemma random mean	0.75	0.1283	—	—	—
Gemma — unsafe-alpha diagnostic (appendix only)					
Gemma targeted (unsafe)	1.0	0.5333	0.3333	85.0	0.1905
Qwen2.5-7B-Instruct, L25 pre_first — causal failure					
Qwen targeted (diagnostic)	**0.75**	**0.1000**	**1.0000**	**0.0**	**1.00**
Qwen random mean	0.75	0.6517	—	—	—
Qwen — unsafe-alpha diagnostic (appendix only)					
Qwen targeted (unsafe)	1.0	0.9667	0.5667	100.0	0.0909

Strict percentiles are retained as historical effect-position displays; $p_{\mathrm{MC}} = (1 + \#\{b : T_b \geq T_{\mathrm{target}}\})/(B + 1)$ is the corrected finite-control value. Dense-family Holm-adjusted p-values are: Gemma $\alpha = 0.75$: 0.3636; Gemma $\alpha = 1.0$: 0.3810; Qwen $\alpha = 0.75$: 1.0000; Qwen $\alpha = 1.0$: 0.3636. Gemma dose-response (α: change rate): 0.25:0.067, 0.50:0.133, 0.625:0.200, 0.75:0.233, 0.875:0.400, 1.0:0.533. Controls at $\alpha = 0.75$: Gemma shuffled-label$= 0.10$ (direction control); direction-only add$= 0.40$; norm-rescaled mean$= 0.233$. Qwen direction-only add$= 0.50 \gg$mean-replace$= 0.10$ (different update rules; norm mediation is not identified). Exploratory layer effects: Gemma L31 0.43 $\gg$ L20 0.10 $\gg$ L9 0.03; Qwen L25 0.97 $\gg$ L7 0.07. Qwen $\alpha = 1.0$ random distribution uses v1 data (10 seeds) due to a v2 process crash; its p_{MC} denominator is therefore 11.

A.4 Causal Diagnostics

This section provides appendix-only historical activation-patching diagnostics for Section 6. The figures retain the original effect estimates, but their old finite-control interpretation is superseded by the corrected plus-one Monte Carlo rule with ties counted against the target:

$$p_{\mathrm{MC}} = \frac{1 + \#\{b : T_b \geq T_{\mathrm{target}}\}}{B + 1}.$$

No dense-intervention row passes Holm correction within the four-row dense family. These diagnostics are exploratory specificity checks, not a deployable mitigation, an operating recommendation, or cross-family causal evidence. Because intervention figures are no longer main-text floats, the historical figure equivalents are retained here with corrected inferential captions.

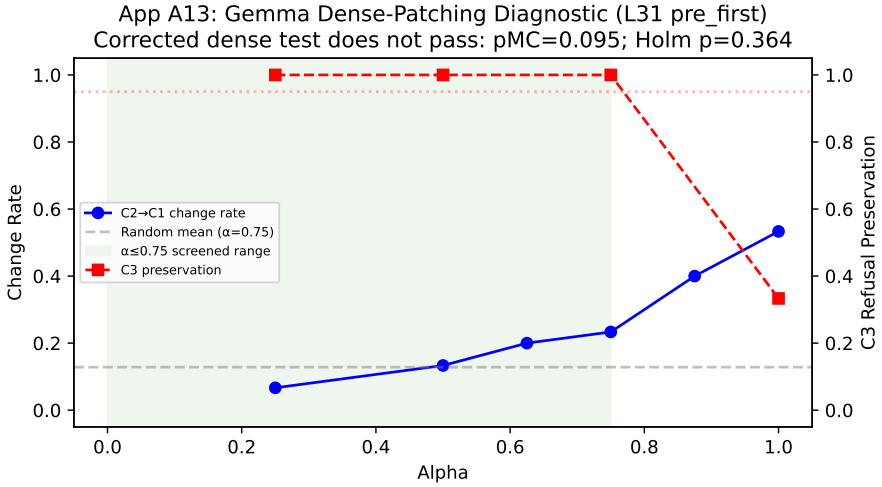


Figure 15: Gemma-2-9B historical alpha frontier. Dose-response of targeted C2→C1 mean-replacement patch at L31 `pre_first` (blue). C3 safety preservation is 1.00 for $\alpha \leq 0.75$ in the tested in-distribution C3 subset and drops to 0.33 at $\alpha = 1.0$. **Gemma only; exploratory diagnostic; safety is cohort-scoped and in-distribution only.**

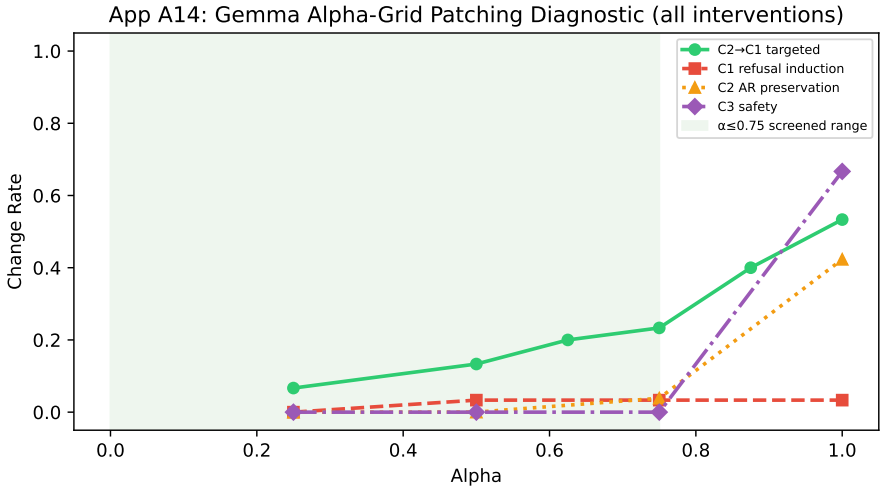


Figure 16: Gemma historical alpha-grid dose response for all four intervention types (targeted C2→C1, C1 refusal induction, C2 AR preservation, C3 safety). $\alpha = 1.0$ **is unsafe; appendix only. Targeted C2→C1 shows monotonic dose-response; C1 refusal induction remains** ≤ 0.033 **across all** α**. This is not corrected random-control evidence.**

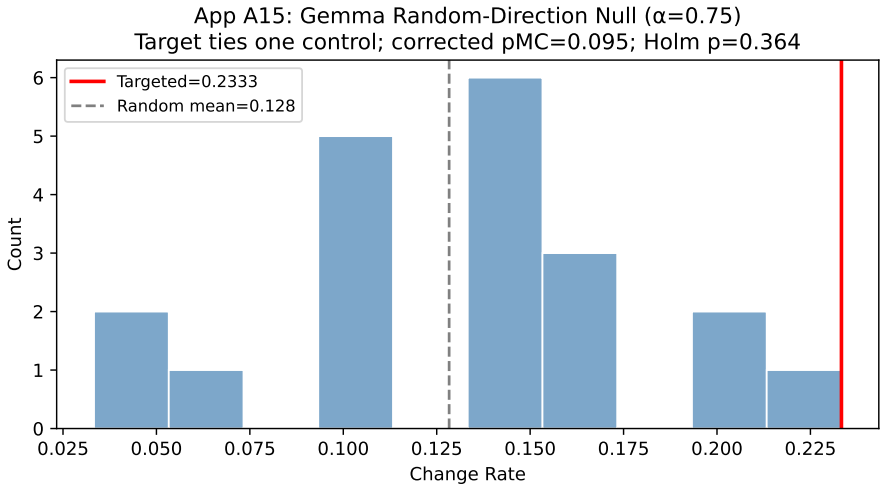


Figure 17: Distribution of 20 random-direction change rates vs the targeted change rate at $\alpha = 0.75$ in Gemma. Targeted $= 0.2333$ (red line). **Historical strict percentile** $= 95.0\%$ **with one random direction tied. Corrected tie-inclusive** $p_{\text{MC}} = 2/21 = 0.0952$ **and dense-family Holm** $p = 0.3636$**; no corrected dense pass.** Shuffled-label control $= 0.10$ (dashed).

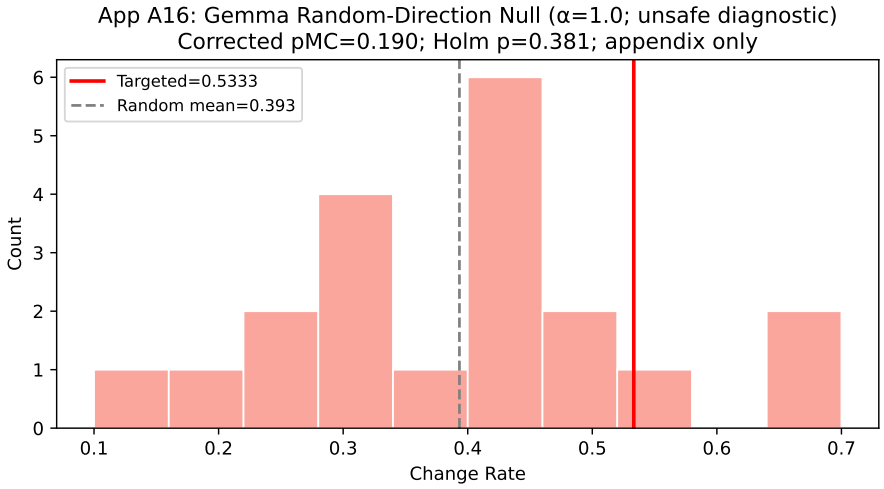


Figure 18: Distribution of 20 random-direction change rates vs the targeted change rate at $\alpha = 1.0$ in Gemma. Targeted $= 0.5333$ (red line). **Unsafe appendix-only diagnostic. Corrected** $p_{\text{MC}} = 4/21 = 0.1905$ **and dense-family Holm** $p = 0.3810$**; C3 preservation drops to 0.33.**

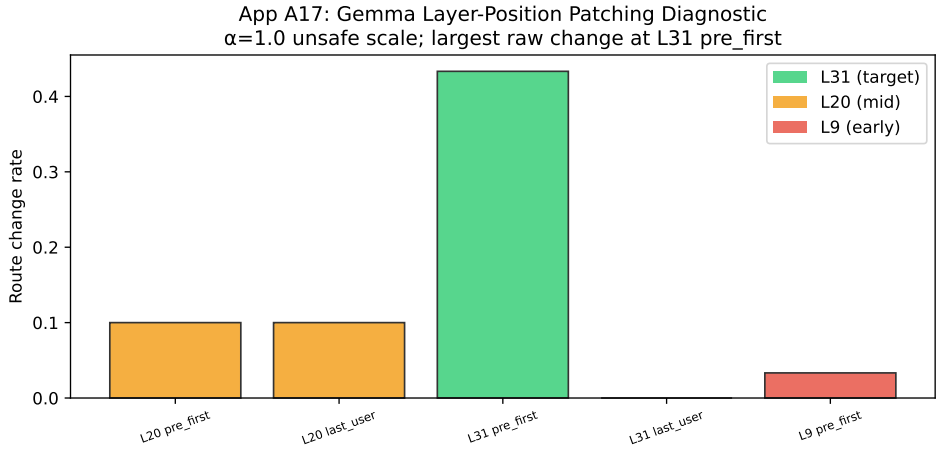


Figure 19: Gemma layer-position sweep at $\alpha = 1.0$. L31 `pre_first` (change rate 0.43) dominates L20 `pre_first` (0.10) and L9 `pre_first` (0.03). L31 `last_user` is inert (0.00). **Unsafe historical setting; layer specificity diagnostic only.**

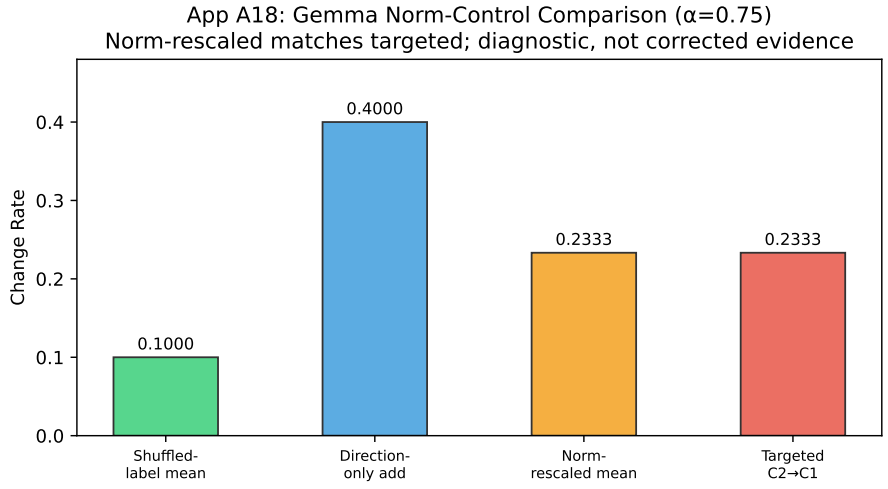


Figure 20: Gemma norm-control comparison at $\alpha = 0.75$. Norm-rescaled mean (0.2333) has the same aggregate rate as targeted C2→C1 (0.2333). Direction-only add gives 0.40 versus mean-replace 0.2333. These results are consistent with a norm-related alternative and do not establish direction specificity. Equal marginal rates do not imply identical changed prompts or identify a norm-mediated mechanism.

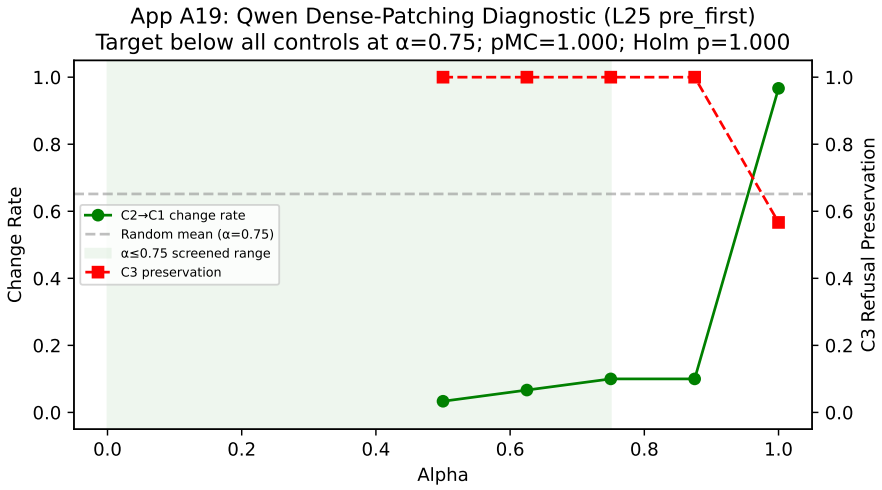


Figure 21: Qwen2.5-7B historical alpha frontier. Targeted C2→C1 change rate (0.10) is below the random-direction mean (0.65) at $\alpha = 0.75$. **Qwen does not replicate the Gemma intervention pattern; Qwen evidence in this paper is predictive-only.**

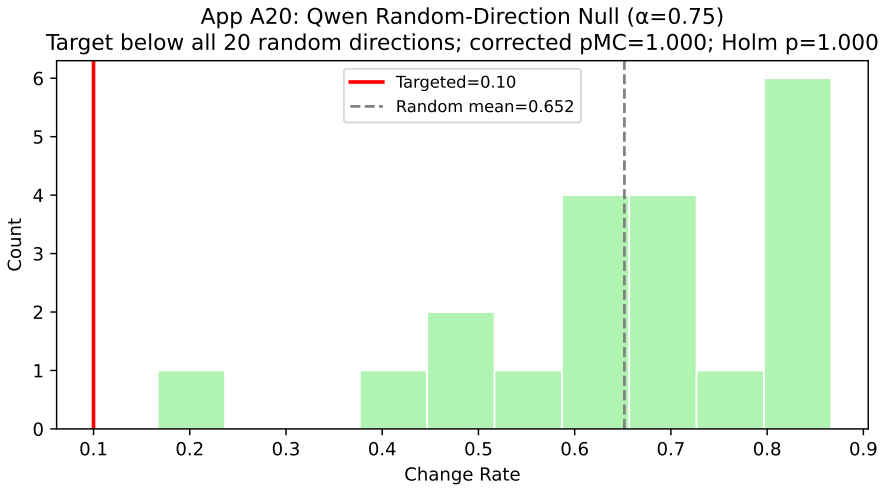


Figure 22: Qwen2.5-7B random-direction distribution at $\alpha = 0.75$. Targeted $= 0.10$ (red line), **below ALL 20 random-direction change rates** ($\min = 0.17$, mean $= 0.65$). Corrected $p_{\mathrm{MC}} = 21/21 = 1.0000$. **No cross-family causal claim; Qwen is predictive-only.**

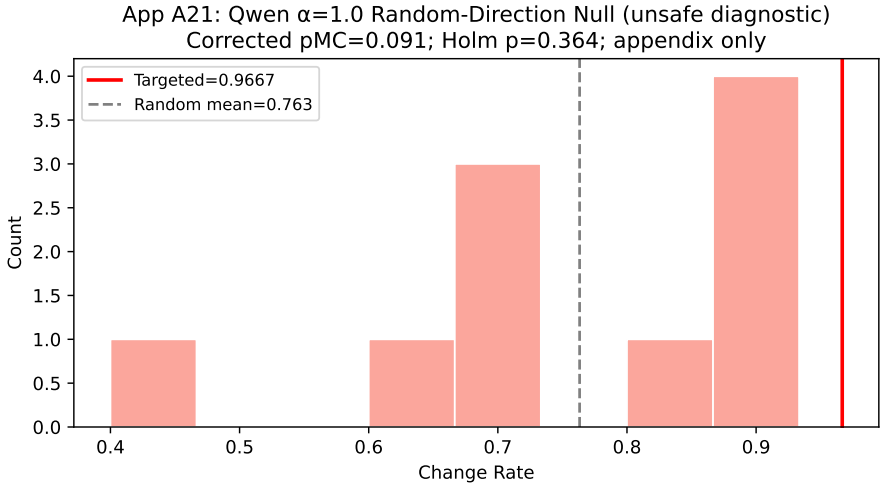


Figure 23: Qwen2.5-7B $\alpha = 1.0$ random-direction distribution. Targeted $= 0.9667$. Uses v1 data (10 seeds) because the v2 process crashed during the $\alpha = 1.0$ random phase. Corrected $p_{\text{MC}} = 1/11 = 0.0909$, but this is an **unsafe appendix-only diagnostic**; C3 preservation drops to 0.57.

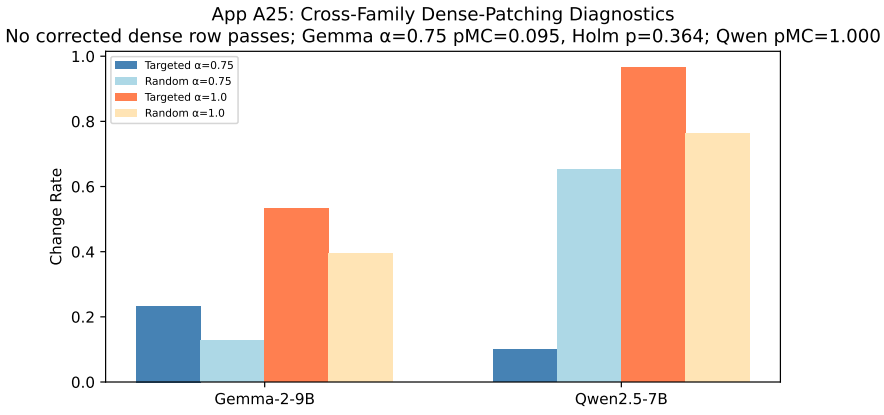


Figure 24: Cross-family causal comparison: targeted vs random at $\alpha = 0.75$ and $\alpha = 1.0$ for Gemma-2-9B and Qwen2.5-7B. **Historical effect display only: corrected dense-family Holm tests do not pass, Qwen does not replicate the Gemma intervention pattern, and $\alpha = 1.0$ is unsafe.**

A.5 SAE Feature Analysis

This section provides the detailed SAE feature analysis that supports the main-text summary in Section 6.1. The historical intervention is retained as a preliminary diagnostic on feature **indices** (top-10 by Cohen's d), not as a confirmed specificity test. The corrected finite-control value for the top-10 downscale is $p_{\mathrm{MC}} = 1/21 = 0.0476$ in a one-row retrospective SAE family, but the historical random null was mostly inactive: the reconstructed random mean already-zero fraction is 0.986, compared with a target active-value fraction of 0.857. Active-feature-matched random-control confirmation is unfinished. We do **not** claim these features constitute a causal circuit, are monosemantic, or reveal the mechanism of false refusal.

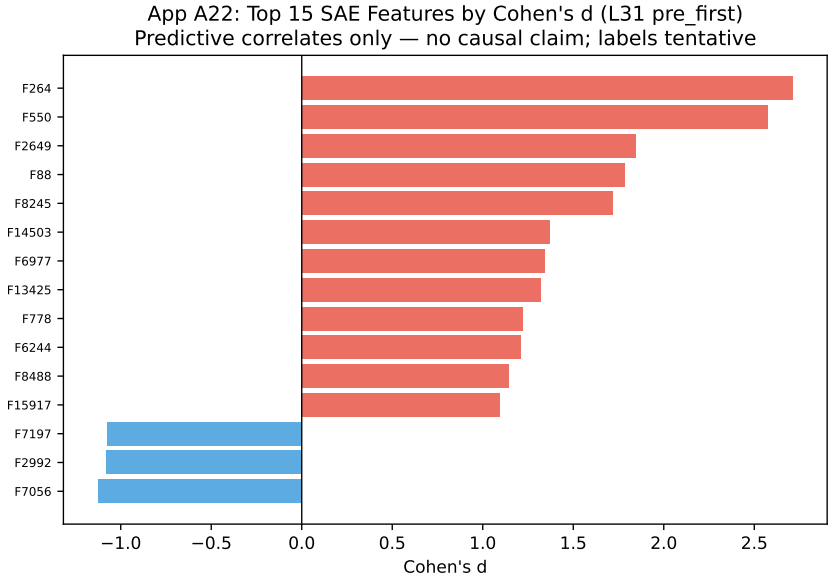


Figure 25: Top 15 SAE features by absolute Cohen's d at layer 31 `pre_first`. Positive direction = more active in false refusal. Feature labels are tentative interpretations from human expert review. **Predictive correlates only; no causal claim; labels tentative; not monosemantic; features may track surface hazard cues rather than refusal decisions.**

Table 8: Historical SAE feature-level intervention in Gemma-2-9B. Gemma Scope SAE, L31, width 16k; $n = 30$ C2 examples per intervention. **Chat template fix applied** (baseline refusal 0.9667; the earlier v2 plain-text baseline of 0.70 was spurious and is superseded). The hazard-cue downscale (50% change, binomial 95% CI $[0.32, 0.68]$) has corrected $p_{\mathrm{MC}} = 1/21 = 0.0476$ within a one-row retrospective SAE family, but the random-feature null was mostly inactive. This is a preliminary diagnostic on feature *indices*, not a causal circuit or completed feature-specificity test. Change counts paired inequality of the historical heuristic route (refusal, answer, or empty) against the regenerated baseline over all 30 prompts, including its one non-refusal. It is not a direct measure of useful-answer recovery.

Feature set	n feat.	Scale	Baseline refusal	Patched refusal	Change rate
Hazard-cue (top 10, Cohen's d)	10	0.0	0.9667	0.4667	**0.5000**
Hazard-cue (dose-response)	10	0.5	0.9667	0.8000	0.1667
Benign-context (top 3, upscale)	3	1.5x	0.9667	0.9667	0.0000
Benign-context (top 3, upscale)	3	2.0x	0.9667	0.9667	0.0000
Random from 16,371-pool (20 sets)	10	0.0	0.9667	—	**0.0000**
Combined hazard-down + benign-up	13	comb.	0.9667	0.9667	0.0000†

†**INVALID**: the combined intervention had a sequential-hook implementation bug—only the second-registered hook (benign-upscale) takes effect during generation, so the reported 0% change is benign-upscale-only, not the true combined effect. The hazard-only downscale (50% change) remains the historical diagnostic effect. **Inertness caveat on the random null:** most of the 16k SAE features are inactive on these inputs, so zeroing them cannot change behaviour; the null is partly trivial. Reconstructed random sets have mean active-value fraction 0.013 and already-zero fraction 0.986 on the v3 subset. An *active-feature* random control (random among features that fire on false-refusal inputs) is the needed stronger null and is not completed here. Benign-context upscaling is inert at both scales in this configuration.

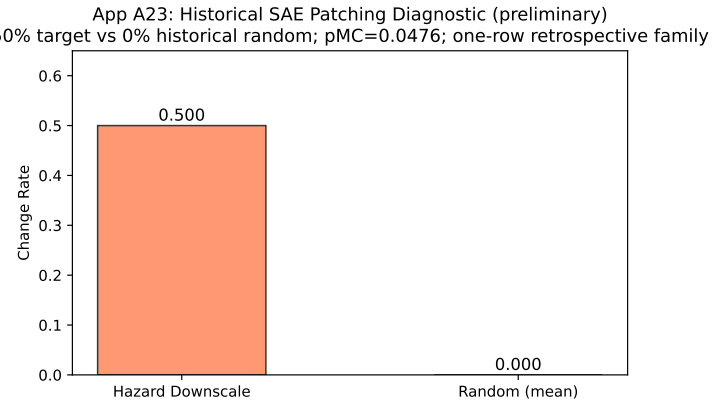


Figure 26: Historical SAE feature-level intervention in Gemma-2-9B (L31, width 16k). Hazard-cue downscaling (scale $= 0.0$) changes 50% of false-refusal routes (binomial 95% CI $[0.32, 0.68]$, $n = 30$), while 20 random-feature sets from the 16,371-feature pool change 0% (strict percentile $= 100.0$, $p_{\mathrm{MC}} = 1/21 = 0.0476$ by the plus-one Monte Carlo rule). **Preliminary diagnostic: not a causal circuit; combined intervention had a hook bug; feature labels are tentative; the historical random null is mostly inactive and active-feature-matched inference is unfinished.**

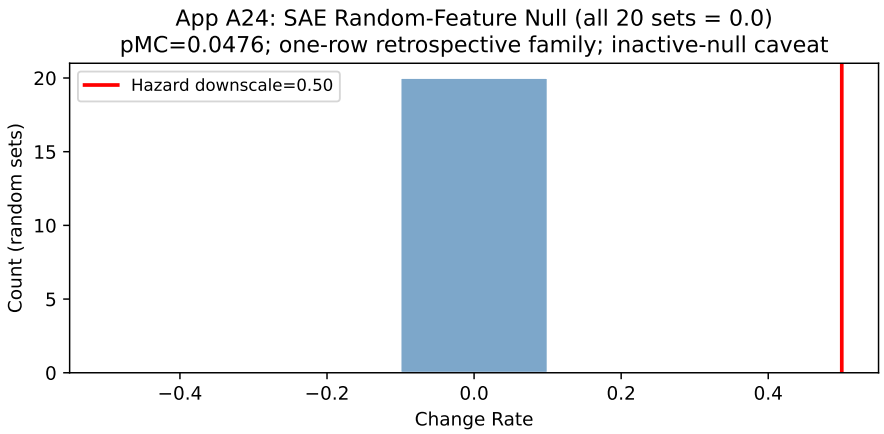


Figure 27: Distribution of change rates across 20 random-feature sets. All 20 produce a 0.0 change rate (strict percentile $= 100.0\%$). **Inactive-null caveat: most randomly sampled SAE features are inactive on these inputs, so zeroing them cannot change behaviour. The random mean active-value fraction is 0.013 and the already-zero fraction is 0.986. This is not a completed feature-specificity test.**

A.6 Labeling and Annotation Provenance Details

This section describes the labeling pipeline and independent human expert annotation methodology. Figures 9 and 10 (Appendix A1) provide the C1 coverage funnel and annotation-audit summaries.

Labeling pipeline. C1/C2/C2-appropriate/ambiguity labels are assigned by human expert annotation of the prompt and the model's response route. C1 (answerable clean) examples have an answerable prompt and a direct or constrained answer. C2 (true false-refusal) examples have an answerable prompt with a `hard_refusal` or `safe_redirection` response. C2 appropriate refusal examples have a refusal-like response to a prompt judged refusal-appropriate (unsafe-or-disallowed) under the prompt audit. C2 ambiguity collapse examples have a refusal-like response to a prompt judged ambiguous. C3 (true-refusal) examples have an unsafe-control prompt with a refusal-like response. C4 examples have an unsafe-control prompt with a non-refusal response. C3/C4 unsafe-control status additionally relies on benchmark design (the prompts are sourced from the unsafe splits of XSTest and OR-Bench-Toxic), so explicit prompt-level labels are partial for unsafe controls. Additional adjudicated prompt labels expand C1 from 257 to 477; the answerability requirement is not relaxed. In the saved expanded mapping, all 477 C1 and 123 C2 records have an answerable prompt label and an audited response route. This full-cohort coverage check is distinct from the 97-row endorsement audit below.

Human expert annotation. Three independent human expert raters audit each sampled example on three axes: prompt label (is the prompt answerable?), response route (hard refusal, safe redirection, constrained answer, direct answer), and final group assignment (C1, C2, C3, other). Adjudication is by majority vote ($\geq$ 2 of 3). This is an **independent human expert audit, human expert annotation**.

Group-OK endorsement statistic. The previously reported 95.9% value is not a kappa-style agreement coefficient. It is the final-group endorsement/pass rate:

$$\frac{\#\{\texttt{final_group_ok_final=yes}\}}{n_{\text{total}}}.$$

The corresponding audit input contains 97 adjudicated rows and separate reviewer-count fields for how many reviewers agreed on prompt, route, and group fields. We therefore describe the statistic as a group-OK endorsement rate, not as kappa, unanimity, or general inter-rater agreement.

Follow-up agent-only cohorts. The historical manuscript results above use the human expert provenance just described. Later confirmatory planning and follow-up pools used Codex-native agent panels and adjudication for prompt-policy or response-route tasks; those agent-only cohorts are not collapsed into the historical human-expert label provenance and are not treated here as completed confirmatory evidence.

Table 9: Historical Gemma dataset and labeling summary. C1/C2/ C2-appropriate/ambiguity groups use *independent human expert-audited* prompt and response-route labels (three-human-expert majority vote; group OK rate 95.9%). C3/C4 *unsafe-control* status comes from benchmark design plus response-route filtering; explicit prompt-level labels are *partial* for unsafe controls (150 of 1,100 prompts have no explicit prompt-level label because their unsafe status is inherent to the benchmark). This is an independent human expert audit. Counts are derived directly from source artifacts. C1 and C2 are Gemma-2-9B response-audited labels. C3 true-refusal counts exclude five `constrained_answer` C3 cases (XSTest) that are not refusal-like routes. OR-Bench-Toxic contributes only unsafe controls (C3/C4); FalseReject, OKTest, OR-Bench-Hard, and PHTest supply no benchmark-design unsafe-control rows here; some prompts can nevertheless be judged refusal-appropriate by the prompt audit. Ambiguity-collapse examples ($n = 13$) are excluded from the analysis subset. **OR-Bench-Hard and OR-Bench-Toxic are shown as separate rows because they play different roles, but they are counted as one OR-Bench benchmark in the five-benchmark summary.**

Dataset	n prompts	C1 clean (neg.)	C2 true false ref. (pos.)	Main subset	C2 app. refusal	C3 true refusal
FalseReject	100	17	23	40	15	0
OKTest	300	207	10	217	0	0
OR-Bench-Hard	100	14	34	48	11	0
OR-Bench-Toxic	50	0	0	0	0	40
PHTest	100	50	22	72	0	0
XSTest	450	189	34	223	0	131
All combined	**1100**	**477**	**123**	**600**	**26**	**171**

C3 true refusal (total = 171) is held out as the safety- preservation subset. Final group OK rate in the separate audit sample: 95.9% (majority endorsement of the existing group assignment, not an inter-rater agreement coefficient). Analysis subset = $477 + 123 = 600$.

A.7 Safety and Release Policy

In-distribution safety only. The C3 safety evaluation is restricted to the in-distribution C3 subset (true-refusal prompts from the same datasets). Out-of-distribution safety is **untested**. At $\alpha = 1.0$, C3 preservation drops to 0.33 (Gemma) and 0.57 (Qwen), demonstrating that the intervention becomes unsafe beyond the evaluated safety-preserving diagnostic setting.

Anonymous aggregate/code supplement. No established gated release is claimed in this submission. The planned anonymous supplement is limited to code, configuration files, manuscript source, aggregate reports, and aggregate/redacted figure or table source data needed to audit the public claims. It does not release raw prompts, raw completions, private annotation artifacts, raw unsafe materials, activation arrays, intervention directions, or label mappings as a public raw-data package. Any broader post-review release would require a separate safety and privacy review.

Dual-use boundary. Intervention artifacts remain dual-use. Historical patches can convert some false-refusal routes to answer-like routes, but high-α settings also degrade true-refusal preservation. We therefore present interventions as diagnostics, not as an operational mitigation or recommendation.

Anonymity. This submission is anonymous. `\iclrfinalcopy` is not enabled. No author names, emails, or institutions appear in any file.

A.8 Reproducibility Details

Nested predictive evaluation. The retrospective nested predictive evaluation uses existing activation matrices and labels only. Outer CV is repeated stratified 5-fold with 10 repeats; candidate selection is performed inside each outer training fold with stratified 4-fold inner CV. Feature selection ranks features by absolute Cohen's d inside the training fold and keeps the top 512; scaling and class-weighted logistic regression fitting are also fold-local. Gemma evaluates dense and SAE representations at layers 9, 20, and 31 and stages `hazard`, `last_user`, `pre_first`, and `post_prefix`. Qwen evaluates dense representations at layers 7, 14, and 25 for `pre_first` only because Qwen SAE and other Qwen stages were not staged. Reported nested intervals are percentile ranges across the ten repeat-level metrics, not independent-sample confidence intervals.

Combined text-plus-metadata baseline. The incremental baseline uses B0 = char-wb TF-IDF text features plus limited metadata; B1 = selected dense only; B2 = text plus metadata plus selected dense; and B3 = text plus metadata plus selected SAE where available. Metadata fields are prompt length, dataset, type, category, and focus; model-output-derived fields are excluded. TF-IDF, metadata encoders, activation feature selection, and scaling are fit inside each outer training fold. Pairwise B2–B0 deltas use source-level mean out-of-fold probabilities, paired source bootstrap intervals with 1,000 samples, and paired swap permutation tests with 1,000 samples.

Finite-control p-values and multiplicity. Finite random-control rows use

$$p_{\mathrm{MC}} = \frac{1 + \#\{b : T_b \geq T_{\mathrm{target}}\}}{B + 1},$$

so ties count against the target. The dense-intervention family has four eligible p-value rows; none passes Holm or BH correction. The historical SAE top-10 downscale row passes only within a one-row retrospective SAE family and remains preliminary because the null controls were mostly inactive.

SAE stability and inactive-null diagnosis. SAE stability and inactive-null analysis uses existing Gemma L31 `pre_first` SAE activations, saved top-feature summaries, saved intervention-result source IDs, and local cached decoder parameters. It runs no model inference, generation, activation capture, dataset download, or checkpoint download. The discovery bootstrap uses 1,000 resamples; mean top-10 Jaccard against the full-sample top 10 is 0.802, with median 0.818. Historical random sets were reconstructed from script seed/rules because exact random feature IDs were not saved in the original v3 tables. On the v3 intervention subset, target active-value fraction is 0.857; reconstructed random mean active-value fraction is 0.013 and already-zero fraction is 0.986.

Alpha = 1.0 activation-patching results. Full change rates, percentile breakdowns, and safety failure for $\alpha = 1.0$ (Gemma C3= 0.33; Qwen C3= 0.57). Reported here as supportive context only; $\alpha = 1.0$ is outside the evaluated safety-preserving diagnostic setting. See Table 7 (unsafe-alpha diagnostic rows).

Qwen alpha = 1.0 random-direction distribution. Uses v1 data (10 seeds) because the v2 process crashed during the $\alpha = 1.0$ random phase.

Direction-only addition at alpha = 0.75. Direction-only addition result at $\alpha = 0.75$ (Gemma 0.40, Qwen 0.50) and the the comparison with mean-replacement. These aggregate rates do not identify norm mediation.

Norm-rescaled mean control. Norm-rescaled mean control and targeted intervention both give Gemma 0.2333 at $\alpha = 0.75$. This is consistent with a norm-related alternative, not proof of a norm-driven mechanism or paired outcome equivalence.

Benign-context upscaling full results. 0% change at 1.5x and 2.0x scale (inert).

Combined intervention hook-bug disclosure. Combined hazard-downscale + benign-upscale sequential-hook bug: only the second-registered hook (benign-upscale) takes effect during generation. The reported combined 0% change is benign-upscale-only and is **INVALID**. See Table 8 note.

OR-Bench-Hard per-dataset CIs. OR-Bench-Hard per-dataset CIs and the text-confound discussion. Treated as a boundary case and excluded from headline claims.

Source data. All source-data CSVs for the figures in this appendix are available in `figures/appendix/source_data/`. No new experiments were run to produce any figure in this appendix.

A.9 LLM USAGE STATEMENT

In accordance with the ICLR policy on large language model (LLM) use, we disclose the role of LLM assistance in this work. GLM-5.2 was used as a research assistant during the project. OpenAI Codex was also used during the current revision for source inspection, LaTeX editing, code/report review, and claim-scope checking. The authors remain responsible for all claims, analysis choices, and submitted text.

Roles in which LLMs were used.

- **Literature retrieval and verification.** Codex assisted with finding related work and checking bibliographic metadata against primary sources during this revision.
- **Code generation and refactoring.** Writing and maintaining the analysis pipeline (probe training, activation extraction, causal-patching, SAE-intervention, and figure-generation scripts).
- **Dataset/schema inspection scripts.** Summarising prompt counts, route distributions, and label-set composition from logged artifacts.
- **Annotation-related tooling.** LLMs helped draft code and reports for handling annotation artifacts. For the historical manuscript cohort, prompt-level, response-route, and feature labels were produced by human experts under majority vote rather than by GLM or Codex.
- **Report drafting and manuscript editing.** Drafting and revising prose, compressing the main text, formatting the LaTeX package, and making the current appendix revision.
- **Co-author-style LLM-assisted review and claim-scope checking.** Role-separated review passes and automated claim-scope checks that flag overclaiming wording (e.g. "causal circuit" used as a claim).

- **Current Codex-assisted review.** Inspecting aggregate reports, checking manuscript/source consistency, and editing LaTeX wording without running new experiments or opening raw private artifacts.

Provenance of reported results. All numerical results reported in the paper come from logged code executions and saved artifacts (CSVs, audit files, and configuration logs)—not from LLM output. No LLM-generated text should be treated as evidence unless it is backed by a saved output, audit file, or source-data CSV.

Annotation provenance. For the historical manuscript cohort, C1/C2/C2-appropriate/ambiguity labels and response-route audits are described as **human expert annotation**; GLM and Codex did not perform those labels. The previously reported 95.9% value is a group-OK endorsement/pass rate, $\#\{\texttt{final_group_ok_final=yes}\}/n_{\text{total}}$, not an inter-rater agreement coefficient. Later follow-up confirmatory pools used Codex-native agent panels/adjudication for prompt-policy or response-route tasks; those agent-only cohorts are reported separately from the historical human-expert manuscript labels and are not treated here as completed confirmatory evidence.

Safety. Unsafe outputs were redacted and are not included in any report or figure. No raw unsafe prompts or completions are released.

Scope of this statement. This statement concerns the role of GLM-5.2 and Codex in the research, review, coding, and writing process. It does not reveal private chain-of-thought or internal prompts. Numerical claims remain tied to saved code executions, aggregate reports, source-data CSVs, and configuration logs.