DECODABLE BUT CONFOUNDED: POKER BLUFF LABELS IN LLAMA-3.1-8B HIDDEN STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

A concept label can be highly decodable from hidden states even when the readout largely exploits task variables that make the label predictable. We make this failure mode directly measurable in poker using 44,631 examples annotated by GPT-4o with binary bluff labels. For Llama-3.1-8B-Instruct, we extract pre-response residual-stream activations at the final prompt token and fit layerwise logistic probes with statement-group-disjoint splits and validation-only layer selection. The label is strongly decodable from both an explicit bluff-classification prompt and a question-stripped neutral-state prompt, reaching held-out AUROC 0.936 and 0.935. However, observed poker-state covariates alone reach AUROC 0.920, and removing activation components linearly predictable from those covariates reduces probe AUROC to 0.722. Because poker state is semantically relevant to bluffing, these controls do not isolate "pure bluff information"; they quantify how much of the headline probe signal overlaps directly observed task structure. Probe-direction interventions produce signed, dose-dependent shifts in normalized next-token $P(\mathrm{Yes})$, but at the strongest tested dose no non-degenerate condition changes the model's argmax Yes/No response, while matched random, shuffled-label, and nuisance directions produce effects of comparable or greater magnitude. A 32,768-feature TopK sparse autoencoder explains 95.2% of held-out activation variance, yet selected single features are weak held-out predictors and learned dictionaries are unstable across seeds. Across probes, interventions, and sparse features, the evidence supports label recoverability and output sensitivity without identifying a bluff-specific mechanism.

1 INTRODUCTION

Linear probes are widely used to ask what information is accessible from neural representations. Their appeal is clear: a simple readout can reveal information that is not obvious from a model's generated text. Their interpretation is less simple. High probe accuracy establishes recoverability of a label, but does not by itself identify which underlying variables support that recovery, whether the signal is specific to the target concept, or whether the decoded direction plays a specific causal role in the model's behavior (Hewitt & Liang, 2019; Belinkov, 2022). This distinction matters whenever a probe is treated as evidence for a concept-specific internal mechanism, especially for safety-relevant concepts such as truthfulness and deception, where ordinary semantic or task variables can strongly predict the target label (Azaria & Mitchell, 2023; Goldowsky-Dill et al., 2025).

We study this problem in a deliberately constrained poker setting. Each example describes a 6-max no-limit Texas Hold'em state and an observed raise that has already occurred. A separate GPT-4o labeling procedure assigns a binary bluff or non-bluff label. We then analyze Llama-3.1-8B-Instruct while it reads the same state and, in the explicit condition, is prompted to classify the observed raise. The subject model is therefore *prompted to judge* an action that has already occurred rather than choosing whether to bluff itself. This distinction is central: our target throughout is the recovered operational bluff label, not deceptive intent or deceptive behavior by the subject model.

Poker makes the interpretability problem unusually transparent because the target label is strongly structured by game-state evidence that is itself semantically relevant to bluffing, including made-hand

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.

category, street, position, and board structure. We therefore do not treat these variables as irrelevant nuisances. Instead, they confound a stronger mechanistic identification claim: a high-performing hidden-state probe may recover the label largely because the model represents the poker state, without isolating a bluff-specific latent variable. Our controls therefore ask how much of the headline probe performance overlaps with directly observed task structure rather than treating game-state predictability as a post hoc caveat.

Our experiments separate three levels of evidence. **Decodability** asks whether the bluff label can be linearly read from pre-response activations. **Causal sensitivity** asks whether intervening along a probe-derived direction changes the model's own output distribution. **Mechanism specificity** asks whether that effect is selective relative to matched control directions and ordinary poker-state directions. We additionally ask whether a sparse dictionary yields stable individual features associated with the label.

Our main findings are:

- **Strong late-layer decodability survives instruction removal.** The explicit bluff-classification prompt reaches held-out AUROC 0.936 at physical layer 30. Removing the explicit bluff question yields AUROC 0.935 at layer 31.
- **Observed poker state accounts for a large fraction of the headline signal.** A poker-state-only model using parsed game-state covariates reaches AUROC 0.920. Train-fitted residualization reduces the primary hidden-state probe from AUROC 0.936 to 0.722, and matching on made-hand category reduces it to 0.843.
- **Probe-direction interventions shift the output probability, but do not establish direction specificity.** Opposite orientations produce opposite, dose-dependent changes in normalized $P(\mathrm{Yes})$. At the strongest tested dose, the argmax response remains "No" for every prompt under all non-degenerate conditions, while matched random, shuffled-label, and made-hand nuisance directions produce effects of comparable or greater magnitude.
- **Sparse reconstruction does not support stable single-feature bluff identification.** A 32,768-feature TopK sparse autoencoder explains 95.2% of held-out activation variance, but the learned dictionaries have low cross-seed stability and selected single features are weak held-out predictors.

The contribution is therefore not the general observation that decodability differs from causality, nor a claim that a high-AUROC probe has discovered a deception representation. Rather, we put several forms of apparently positive evidence on the same frozen experimental substrate: high probe AUROC, persistence after removing the target question, dose-responsive live-model steering, and high-quality sparse reconstruction. We then test each against the controls needed for a stronger mechanistic interpretation. Directly measured poker-state covariates nearly reproduce the headline label predictability, matching and residualization substantially reduce it, control directions can move the same output endpoint as much as or more than the probe direction, and selected sparse features are unstable and weak held-out predictors. The result is a concrete demonstration that an apparently coherent "concept representation" story can survive multiple headline analyses while failing specificity tests.

2 Related Work

Probing internal representations. Linear and lightweight probes are common tools for measuring information accessible from neural activations. Prior work has emphasized that probe performance can reflect properties of the probe, dataset, or correlated variables rather than a uniquely identified internal representation (Hewitt & Liang, 2019; Belinkov, 2022). Work on truthfulness and latent knowledge has shown that model activations can support strong linear readouts in a variety of settings (Burns et al., 2022; Azaria & Mitchell, 2023; Bao et al., 2025). Most directly related, Goldowsky-Dill et al. (2025) study linear probes for strategic deception in settings where models themselves engage in deceptive strategies. Recent work also demonstrates that strong decodability can coexist with weak or nonspecific behavioral effects (Sudheendra & Srivastava, 2026; Tiwari et al., 2026). Our setting is deliberately narrower: the subject model is prompted to classify an already-observed poker action, and the supervised target is an external annotation. This makes a different question unusually

tractable: how much of the apparent concept signal overlaps with directly observed, semantically meaningful task-state variables before a bluff-specific mechanism is inferred.

Activation interventions. Representation engineering and activation steering modify model behavior by adding or replacing activation-space directions (Zou et al., 2023; Li et al., 2023; Turner et al., 2023). Causal interpretability work more broadly stresses that correlational readouts and causal interventions answer different questions (Geiger et al., 2021). We follow this distinction by evaluating the live model's own next-token Yes/No probability distribution, not the probe score, and by comparing the probe direction with matched random, orthogonal, shuffled-label, wrong-position, wrong-layer, patching, and nuisance-direction controls.

Sparse autoencoders. Sparse autoencoders have been proposed as a way to decompose dense language-model activations into more interpretable features (Huben et al., 2024; Gao et al., 2025). Cross-seed instability is already documented, including for TopK SAEs (Paulo & Belrose, 2026). We therefore treat stability as a required diagnostic rather than a novel claim: we report held-out reconstruction, sparsity, dead-feature rate, cross-seed dictionary stability, and held-out single-feature predictiveness, and avoid assigning semantic names to features based only on label association.

Poker benchmarks. The source examples are attributable to PokerBench (Zhuang et al., 2025), a benchmark for poker reasoning and decision making. Our recovered dataset is a derived set of raise examples with GPT-4o bluff judgments. Because exact upstream split and hand identifiers are unavailable, we make no claim that the 44,311 statement-derived groups correspond to independently dealt hands.

3 Data and Experimental Protocol

3.1 Poker bluff-label dataset

The analyzed dataset contains 44,631 examples, of which 7,490 are labeled bluff and 37,141 non-bluff, giving a 16.8% positive rate. Every row describes a 6-max no-limit Texas Hold'em state containing player position, private cards, public board cards, betting context, and an observed raise. The binary target is a GPT-4o yes/no judgment produced by a recovered poker-analysis prompt. We therefore treat the labels as *model-generated annotations*, not game-theoretic ground truth and not measurements of the Llama subject model's intent.

The source examples are attributable to PokerBench, but repository history does not recover the exact PokerBench split or revision, original hand/session identifiers, the final 44,631-row selection script, the upstream procedure that generated the observed actions, or the exact served GPT-4o snapshot. We use statement-derived group identifiers to keep identical statement groups from crossing partitions. These identifiers must not be interpreted as verified independent hands or sessions.

The original explicit prompt asks whether the observed raise is a bluff and requests a yes/no answer. Its response field is the label verbatim. Any activation analysis that reads the response token would therefore be trivially leaked. Our primary extractor never receives the response text. We additionally construct a `neutral_state` variant that removes the explicit bluff question while retaining the same poker state and observed raise. This directly tests whether decodability depends on naming the target concept in the prompt.

3.2 Frozen Statement-Group-Disjoint Split

We use one frozen grouped and label-stratified split shared by downstream analyses: 31,241 training rows, 4,463 validation rows, and 8,927 test rows. Statement-derived groups never cross partitions. Model, layer, feature, and direction choices are made using the training and validation partitions only. The held-out test partition is used after those choices are fixed. Dataset and split checksums are recorded in run artifacts so downstream stages reject stale or mismatched inputs.

3.3 ACTIVATION EXTRACTION AND PROBES

We use `meta-llama/Llama-3.1-8B-Instruct`, with model and tokenizer revision `0e9e39f249a16976918f6564b8830bc894c89659`, loaded in bfloat16. For each example, we record the residual-stream output of all 32 transformer blocks at the final prompt token before answer generation. The hidden dimension is 4096. The explicit variant is read at the prompt boundary ending in `Answer:`; the neutral variant is read at the end of the state-action description. Extraction metadata records prompt hashes, token positions, physical layer mappings, model/tokenizer revisions, activation shapes, and tensor hashes.

At physical layer ℓ, we fit an L2-regularized logistic probe

$$p(y = 1 \mid h_\ell) = \sigma(w_\ell^\top h_\ell + b_\ell),$$

with $C = 1$, class balancing, and the fixed training partition. The selected physical layer maximizes validation AUROC. We then evaluate that frozen selection on test. Five configured probe seeds yield the same fitted validation curves and selected layers in this deterministic setting, so we do not interpret their near-zero spread as independent sampling uncertainty. Sampling variability is instead assessed with grouped bootstrap intervals and independent group-level subsamples in the learning-curve experiment.

We compare against majority accuracy, TF-IDF prompt text, a response-text leakage diagnostic, and a poker-state-only model fitted to parsed poker-state covariates. Because only 16.8% of examples are positive, we record PR-AUC and calibration metrics in addition to AUROC, although AUROC is used for layer selection for consistency with the prespecified protocol.

3.4 POKER-STATE CONTROLS

The parsed covariates include action, street, position, made-hand category, board texture, board pairing, and amount faced. We use "nuisance" as shorthand for these observed task-state covariates in the control analysis, not to imply that they are semantically irrelevant to bluffing. We use three complementary controls. First, a poker-state-only classifier estimates how predictive these observed variables are without hidden states. Second, matched-subset analyses balance labels within strata defined by individual task-state variables. Third, train-fitted residualization removes activation components linearly predictable from the observed covariates before refitting the bluff probe.

These controls do not establish that the residualized representation is "pure bluff information." The covariate set is incomplete, residualization is linear, and game state is semantically related to bluffing. We therefore interpret the controls as measurements of overlap with observed task structure, not as a decomposition of the representation into "true signal" and "confound." True hand equity is unavailable, and reliable pot size cannot be reconstructed from the recovered text, so equity-matched and pot-size-matched controls are explicitly recorded as not run.

3.5 CONTROLLED LIVE-MODEL INTERVENTIONS

For the explicit prompt, we fit a probe-derived direction using the training partition and intervene in the live Llama model at the validation-selected physical layer. The primary endpoint is the model's own forced-choice output distribution over the tokens "Yes" and "No", never the probe score. The intervention subset contains 500 held-out examples spanning 495 statement-derived groups.

For a hidden state h and unit-norm direction d, steering uses

$$h' = h + \alpha d,$$

with $\alpha \in \{0.5, 1, 2, 4\}$ and both orientations of d. These doses correspond to mean perturbation magnitudes of approximately 1.08%, 2.16%, 4.32%, and 8.65% of the hidden-state norm in the evaluated examples. We compare the target direction with matched-norm random, orthogonal, shuffled-label, wrong-layer, wrong-token-position, mismatched-prompt activation-patching, and made-hand nuisance controls. Effects are paired within prompt and summarized with group-aware bootstrap confidence intervals. We also track probability mass assigned to the allowed yes/no choices to distinguish targeted endpoint movement from generic model degradation.

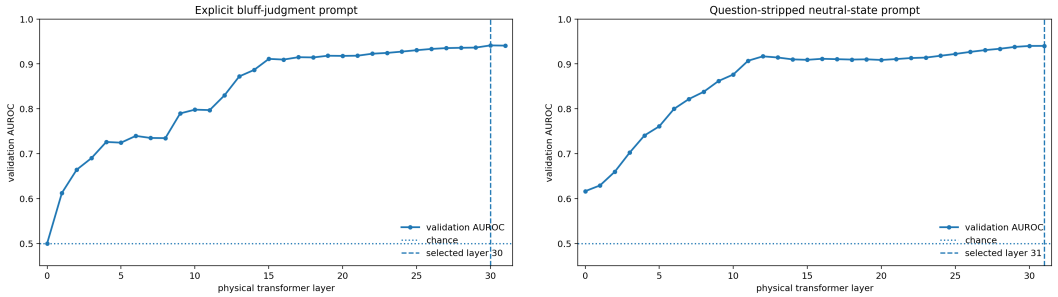

Figure 1: Layerwise validation AUROC for the explicit bluff-classification prompt (left) and question-stripped neutral-state prompt (right). Layer selection uses validation only. Because the configured logistic-probe seeds produce the same deterministic fits, no seed-variance band is shown; group-level learning-curve resampling in Appendix C measures sampling variability.

Table 1: Headline held-out results. Grouped-bootstrap 95% confidence intervals are shown where available; "–" indicates that a compatible interval was not recorded for that analysis.

Condition	Layer	Test AUROC	95% CI
Explicit bluff classification	30	0.9360	[0.9300, 0.9418]
Neutral state	31	0.9351	[0.9287, 0.9414]
Poker-state metadata only	–	0.9204	–
Residualized explicit probe	30	0.7220	–

3.6 Sparse autoencoder

At the explicit prompt's validation-selected layer, we train an unsupervised TopK sparse autoencoder with 32,768 features, an $8\times$ expansion over the 4096-dimensional residual stream, and exactly 64 active features per example. Training uses 60 epochs, batch size 1024, learning rate 3×10^{-4}, train-only input normalization, and unit-norm decoder columns. Labels are not used to fit the dictionary. After training, the dictionary is frozen, features are ranked using training plus validation data, and selected features are scored once on the test set. Three training seeds are used for dictionary-stability analysis.

4 Results

4.1 Bluff labels are strongly decodable in late layers

The explicit bluff-classification prompt selects physical layer 30 and reaches validation AUROC 0.9413 and held-out test AUROC 0.9360. The neutral-state variant selects physical layer 31 and reaches validation AUROC 0.9402 and test AUROC 0.9351. Thus removing the question asking whether the raise is a bluff changes test AUROC by less than 0.001.

Figure 1 shows a broad increase in linear decodability toward later layers. The neutral condition is especially informative because its final token occurs before any request to classify bluffing. Strong late-layer separability therefore does not depend on the model being explicitly asked the target question. It still does not imply a bluff-specific representation, because both variants encode the same poker state.

The prompt-text-only baseline reaches AUROC 0.7817 for the explicit prompt and 0.7825 for the neutral prompt, well below the hidden-state probes. Group-level learning curves also increase smoothly with data: for the explicit condition, mean test AUROC rises from 0.762 at 500 training groups to 0.936 at 31,000; the neutral condition rises from 0.710 to 0.935 over the same sizes (Appendix C).

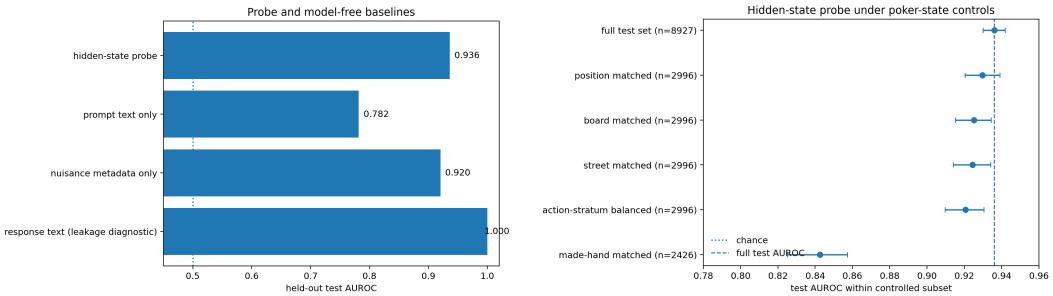

Figure 2: Poker-state controls for the explicit prompt. Left: held-out AUROC for the hidden-state probe and baselines. The response-text result is a leakage diagnostic, not evidence. Right: hidden-state probe performance under controlled subsets with grouped-bootstrap 95% intervals; the action-stratum-balanced row is not evidence of a remaining action-type confound because all rows are raises.

4.2 Observed poker state explains a large fraction of the signal

The poker-state-only classifier reaches test AUROC 0.9204 and PR-AUC 0.6166, approaching the explicit hidden-state probe's AUROC of 0.9360 despite using no model activations. The source labels are strongly associated with made-hand category (Appendix A), and the same task-state variables are represented in the model. Thus the raw probe score alone cannot distinguish a bluff-specific representation from state-mediated label prediction.

Train-fitted residualization provides a stronger control. Removing activation components linearly predictable from the observed nuisance covariates reduces test AUROC from 0.9360 to 0.7220, an absolute decrease of 0.2141. Matched-subset results in Figure 2 point in the same direction. Matching on position yields AUROC 0.9297, board structure 0.9251, street 0.9244, and made-hand category 0.8426. Because every analyzed row already describes a raise, the action-stratum-balanced result (AUROC 0.9207) mainly reflects label balancing within a single action category and should not be interpreted as evidence that action identity is an additional remaining confound.

The late hidden states also make the nuisance variables themselves highly decodable: street is predicted with 1.000 accuracy, position with approximately 1.000 accuracy, board texture with 0.914, board pairing with 0.966, and made-hand category with 0.756; amount faced is predicted with $R^2 \approx 0.966$. These diagnostics are consistent with a representation that contains rich poker-state information from which the operational bluff label can be inferred.

The residualized AUROC of 0.722 remains above chance, but it is not an estimate of an isolated bluff variable. The observed covariates are incomplete, true equity and reliable pot size are unavailable, and linear residualization need not remove nonlinear or unobserved state information. Because these covariates are also semantically relevant to the label, the supported conclusion is deliberately narrower: a large portion of the raw probe signal overlaps directly observed poker-state information, and the remainder is not identified as bluff-specific.

4.3 Interventions shift the output-probability endpoint without establishing direction specificity

Intervening along opposite orientations of the learned probe direction produces opposite and approximately dose-dependent changes in the model's normalized next-token $P(\text{Yes})$ within the allowed Yes/No pair. At the largest tested strength, corresponding to a mean perturbation magnitude of 8.65% of the hidden-state norm, adding and subtracting the probe direction shift this probability by -0.0192 (95% CI $[-0.0204, -0.0179]$) and $+0.0177$ (95% CI $[0.0163, 0.0192]$), respectively. Importantly, the baseline argmax response is "No" on all 500 prompts, and at the strongest tested dose it remains "No" on all 500 prompts for every non-degenerate intervention condition shown in the quality-control analysis. The reported intervention effects are therefore probability shifts, not discrete decision flips.

The control suite prevents a stronger interpretation. At the same target norm, a matched random direction changes the endpoint by $+0.0470$, a shuffled-label direction by $+0.0272$, an orthogonal direction

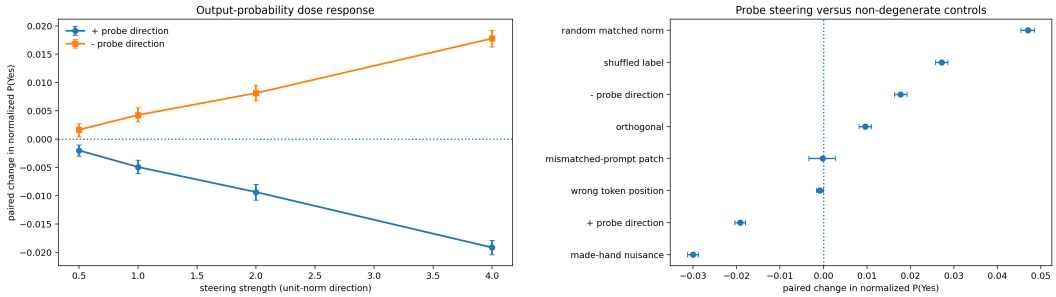

Figure 3: Controlled live-model interventions on 500 held-out prompts. Left: paired change in normalized $P(\text{Yes})$ across probe-direction doses. Right: strongest tested effect for each non-degenerate intervention condition; the degenerate wrong-layer result is reported separately in Appendix D. Several controls have effects comparable to or larger than the probe direction.

by $+0.0096$, and a made-hand nuisance direction by -0.0300. The wrong-token-position control is near zero (-0.0008), and mismatched-prompt activation patching is also near zero (-0.0002). A wrong-layer intervention yields a much larger apparent effect, but it is degenerate: valid yes/no probability mass collapses from approximately 0.978 at baseline to 0.040, while the positive-choice rate becomes 0.992. We therefore treat that point as output degradation rather than targeted causal control.

For the intended probe interventions and the non-degenerate control directions, valid-choice mass remains near 0.976–0.980. The experiment therefore shows causal sensitivity of the model's next-token Yes/No probability distribution to perturbations along the learned direction, but it does *not* show a discrete judgment change or that this direction uniquely identifies a bluff mechanism. In this setting, causal output sensitivity and mechanism specificity are empirically distinct.

4.4 Sparse reconstruction is strong, but individual features are weak and unstable

The TopK sparse autoencoder explains 95.17% of test activation variance while maintaining mean L_0 of approximately 64 active features per example. However, 75.6% of dictionary features are dead on the test partition. More importantly for semantic interpretation, cross-seed dictionary stability is low: the mean maximum cosine similarity across seed pairs is approximately 0.101. The ten features selected using training plus validation label separability achieve held-out single-feature AUROCs spanning approximately 0.42 to 0.59, with the best at 0.593.

These results separate reconstruction from semantic identification. Consistent with prior reports of seed-dependent SAE dictionaries (Paulo & Belrose, 2026), this dictionary provides an accurate sparse reconstruction of the selected activation space but does not support stable, monosemantic "bluff features." We do not interpret this as evidence that other SAE architectures or training choices would necessarily fail. Full SAE diagnostics are shown in Appendix E.

5 Discussion

The central result is a failure of specificity, not a failure of decodability. An operational bluff label is strongly linearly recoverable from late Llama-3.1-8B hidden states, ordinary poker-state covariates nearly match that probe and account for a large portion of its performance under matching and residualization, and a probe-derived direction can move the model's normalized Yes/No output probabilities. Yet several control directions move the same endpoint by comparable or greater amounts, and the non-degenerate interventions do not flip the argmax response at the strongest tested dose. The evidence therefore supports recoverability and interventional output sensitivity while withholding the stronger claim that a bluff-specific mechanism has been identified.

This combination matters for how activation-level results are interpreted. A high-performing probe answers a recoverability question; it does not identify why the information is recoverable. In

structured domains, a model may encode task variables that are strongly predictive of the target without isolating a concept-specific mechanism. The neutral-state condition rules out one simple explanation, namely that the probe merely reflects an explicit instruction to classify bluffing, but the poker-state controls reveal a deeper source of predictability. Because those state variables are part of the evidence relevant to bluffing, our claim is not that they are spurious. The claim is that raw decodability does not distinguish state-mediated label prediction from a more specific bluff representation.

The intervention experiment raises a parallel distinction. Moving a model-output endpoint after perturbing a direction is stronger evidence than moving the score of the probe that defined that direction. Even so, endpoint movement does not establish target specificity or a changed discrete decision. Here, random, shuffled-label, and nuisance directions can move normalized $P(\mathrm{Yes})$ as much as or more than the probe direction, while the non-degenerate argmax remains unchanged. A mechanism-specific claim would require the target intervention to separate clearly from these controls, ideally across additional models, tasks, and model-as-actor settings.

The SAE result provides a third version of the same lesson. Strong reconstruction quality does not imply that individual sparse coordinates correspond to stable semantic units. Low dictionary stability, high dead-feature rate, and weak held-out single-feature discrimination make strong feature naming unwarranted in the present setup. This is a limitation of the evidence from the trained dictionaries, not a general negative result about sparse autoencoders.

Taken together, the experiments suggest a practical evidence hierarchy for concept probes in structured tasks: compare hidden-state readouts with directly observed task-state baselines; test how much performance survives matching or train-fitted removal of those variables; evaluate interventions against matched control directions and output-quality diagnostics; and require held-out stability before assigning semantics to sparse features. Our results do not prove that this sequence is sufficient for mechanism identification, but they show why skipping these steps can make a non-specific signal look mechanistically compelling.

6 LIMITATIONS

Operational labels. The target labels are GPT-4o judgments rather than human or solver adjudications. Their validity is inherited from the recovered labeling prompt, and the exact served GPT-4o snapshot is unavailable. Future work should compare model annotations with independently adjudicated bluff definitions, especially for draws and semi-bluffs.

Dataset lineage and independence. The source is attributable to PokerBench, but the exact upstream split/revision, hand/session identifiers, final selection script, and action-generation procedure are not recoverable. Statement-derived grouping protects against direct statement-group leakage across partitions but does not prove independence across underlying poker hands.

Incomplete strategic controls. We cannot run true equity matching or reliable pot-size matching with the recovered metadata. Residualization is linear and limited to observed covariates. Consequently, residualized performance should not be interpreted as an estimate of "pure" bluff information.

One subject model and one task context. The strict experiments reported here use Llama-3.1-8B-Instruct and one poker classification task. Historical runs on other architectures are not used as final evidence because they were not repeated under the strict protocol. Cross-context or cross-architecture generalization under the final protocol is therefore not claimed.

Annotation target versus subject-model response. The probe target is the GPT-4o annotation, whereas the intervention endpoint is Llama's own forced-choice Yes/No distribution. These analyses therefore answer different questions: whether the external annotation is recoverable from Llama's hidden state, and whether perturbing that state changes Llama's output distribution. The intervention experiment was designed as a sensitivity test, not as a validation that Llama behaviorally agrees with the annotation source.

Output probability, not a decision flip. At baseline the subject model's argmax forced-choice response is "No" for all 500 intervention prompts. At the strongest tested dose, every non-degenerate intervention condition retains the same argmax response on all 500 prompts. The causal result therefore concerns changes in the normalized Yes/No probability distribution, not observed changes in discrete bluff judgments.

Finite intervention controls. The random matched-norm, orthogonal, and shuffled-label controls each use a single sampled direction or permutation. They show that the observed probe-direction effect is not uniquely large relative to these sampled controls, but they do not characterize the full distribution of effects from unrelated directions.

Judgment, not deceptive behavior. The subject model evaluates an observed raise rather than choosing a poker action. The intervention changes a model-output probability endpoint, not the model's propensity to deceive. A model-as-actor experiment would be required to make claims about deceptive behavior.

7 Conclusion

Poker bluff labels are strongly decodable from late Llama-3.1-8B hidden states, and this result persists after removing the explicit bluff question. On its own, that result could suggest a concept-specific representation. The broader evidence does not support that inference: observed poker-state variables approach the raw probe's predictability, matching and residualization reveal substantial overlap with ordinary task structure, control directions can shift the same output endpoint as much as or more than the probe direction, and sparse reconstruction does not yield stable, strongly label-predictive individual features. Thus this case study supports label decodability and interventional output sensitivity without identifying a bluff-specific mechanism. More generally, it illustrates why decodability, output sensitivity, and mechanism specificity should be treated as separate evidential claims rather than successive conclusions.

Ethics Statement

We did not collect human-participant data for this study. The work concerns white-box analysis of model representations and a strategic-game classification task. Activation monitoring may support model oversight, while activation steering can also manipulate model outputs. We therefore avoid presenting the reported probe as a deployment-ready deception detector and emphasize controls that limit stronger mechanistic interpretations.

Reproducibility Statement

Appendix G records the pinned model/tokenizer revision, split construction, activation read positions, probe and SAE hyperparameters, intervention settings, and execution sequence. The strict pipeline records dataset and split checksums, prompt hashes, activation hashes, environment metadata, and figure-source manifests. Publication figures are generated from recorded result artifacts and have corresponding source-data tables. Controls requiring unavailable metadata are represented explicitly as not run rather than silently approximated.

AI Use Statement

Generative AI tools were used to assist with grammar, sentence structure, readability, and critical feedback on the presentation and framing of existing results. They were not used to generate data, run experiments, implement the experimental pipeline, or generate figures. All experimental results and final scientific claims were reviewed and approved by the authors, who take responsibility for the manuscript.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.

Yuntai Bao, Xuhong Zhang, Tianyu Du, Xinkui Zhao, Zhengwen Feng, Hao Peng, and Jianwei Yin. Probing the geometry of truth: Consistency and generalization of truth directions in LLMs across logical transformations and question answering tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 682–700. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.38.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception with linear probes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19755–19786. PMLR, 2025. URL `https://proceedings.mlr.press/v267/goldowsky-dill25a.html`.

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*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *International Conference on Learning Representations*, 2026.

Smitha Muthya Sudheendra and Jaideep Srivastava. When decodability is not enough: Logical validity representations, behavioral dissociation, and causal tests in language models. *arXiv preprint arXiv:2609.02438*, 2026.

Devesh Tiwari, Camille Davis, Shivank Sinha, Talia Weaver, Aditya Shah, and Maheep Chaudhary. Decodability is not causality: Dissociating probe readouts from behavioral drivers via sae decomposition. *arXiv preprint arXiv:2609.18080*, 2026.

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

Richard Zhuang, Akshat Gupta, Richard Yang, Aniket Rahane, Zhengyu Li, and Gopala Anumanchipalli. PokerBench: Training large language models to become professional poker players. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26175–26182, 2025. doi: 10.1609/aaai.v39i24.34814.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Dataset Composition and Label Structure

The dataset contains 44,631 rows and 44,311 statement-derived groups. The frozen split contains 31,241 training rows, 4,463 validation rows, and 8,927 test rows. The positive rate is 16.8% in the full dataset.

All analyzed rows contain the same observed action category, a raise. Made-hand category is strongly associated with the GPT-4o bluff label in the recovered dataset:

Table 2: Observed bluff-label rate by made-hand category in the recovered source data. These descriptive rates motivate the confound controls; they are not held-out model results.

Made hand	Bluff-label rate	n
Quads	0.000	400
Trips	0.006	7,541
Full house	0.008	3,706
Two pair	0.010	8,029
Pair	0.094	12,216
Straight flush	0.200	105
Flush	0.354	4,576
Straight	0.455	4,896
High card	0.734	3,162

A response-text classifier reaches AUROC 1.0 by construction because the response field is the yes/no label verbatim. We include this only as a leakage diagnostic. The primary activation extractor operates before the response and never receives this field.

B Prompt Variants and Read Positions

The explicit variant retains the original bluff-classification question and appends an `Answer:` cue. Activations are recorded at the final prompt token immediately before the answer. The neutral variant removes the target question and ends at the observed action description, whose final marker is `Raise.` Both variants use the same underlying poker examples, split policy, subject-model revision, and full 32-layer extraction.

The extraction configuration sets a maximum sequence length of 512 tokens. The longest observed prompt is approximately 230 estimated tokens, leaving headroom so truncation does not move the decision boundary. The strict extractor records each rendered prompt hash and token index and rejects provenance-inconsistent artifacts.

C Learning Curves

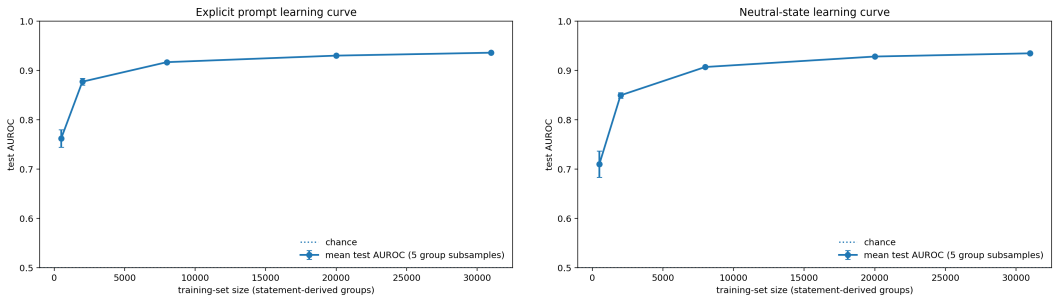

Figure 4: Test AUROC as a function of training-set size for the explicit condition (left) and neutral condition (right). Each point summarizes five independent statement-group subsamples. The same frozen test partition is used throughout.

For the explicit condition, mean test AUROC across the five group-level subsamples is 0.7624, 0.8772, 0.9168, 0.9300, and 0.9359 at 500, 2,000, 8,000, 20,000, and 31,000 training groups, respectively. The corresponding neutral-state values are 0.7100, 0.8495, 0.9069, 0.9281, and 0.9346.

D Intervention Quality Controls

Baseline valid-choice probability mass over the allowed yes/no options is approximately 0.978. For both probe-direction orientations, random matched-norm steering, shuffled-label steering, orthogonal steering, made-hand nuisance steering, wrong-token-position steering, and mismatched-prompt patching, mean valid-choice mass remains approximately 0.976 to 0.980.

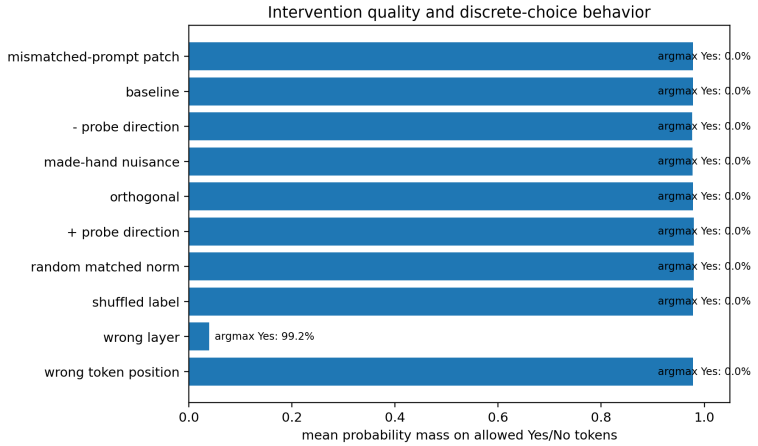

Figure 5: Intervention quality at the strongest reported setting. Bars show mean probability mass on the allowed Yes/No tokens; annotations report the argmax Yes rate. The baseline and every non-degenerate condition have 0% argmax Yes responses, whereas the wrong-layer condition is degenerate: valid-choice mass collapses to approximately 0.040 and the argmax Yes rate rises to 99.2%.

The intervention strength grid $\alpha = \{0.5, 1, 2, 4\}$ corresponds to mean relative magnitudes of 0.0108, 0.0216, 0.0432, and 0.0865 of the hidden-state norm. The largest intended intervention therefore changes the hidden state by about 8.65% of its norm on average.

E Sparse Autoencoder Diagnostics

The SAE has 32,768 features with TopK $k = 64$. On test, reconstruction MSE is 0.0504, fraction of variance explained is 0.9517, mean L_0 is approximately 64, and the dead-feature fraction is 0.7558. Cross-seed dictionary stability, measured by mean maximum cosine similarity across seed pairs, is approximately 0.1009. The ten selected features have test AUROCs ranging from approximately 0.420 to 0.593.

F Additional Confound Details

The neutral-state hidden-state probe reaches AUROC 0.9351. On the same split, poker-state metadata reaches 0.9204. In neutral-state controlled subsets, position matching gives AUROC 0.9285, street matching 0.9270, board matching 0.9192, the single-action label-balanced subset gives 0.9184, and made-hand matching gives 0.8216.

True hand equity is not available in the recovered dataset. Made-hand category should not be treated as a substitute for equity because draws can have high equity while still appearing as "high card." Reliable pot size is also unavailable because reconstructing it would require a complete action-sequence replay. The corresponding equity-matched and pot-size-matched analyses are therefore not run.

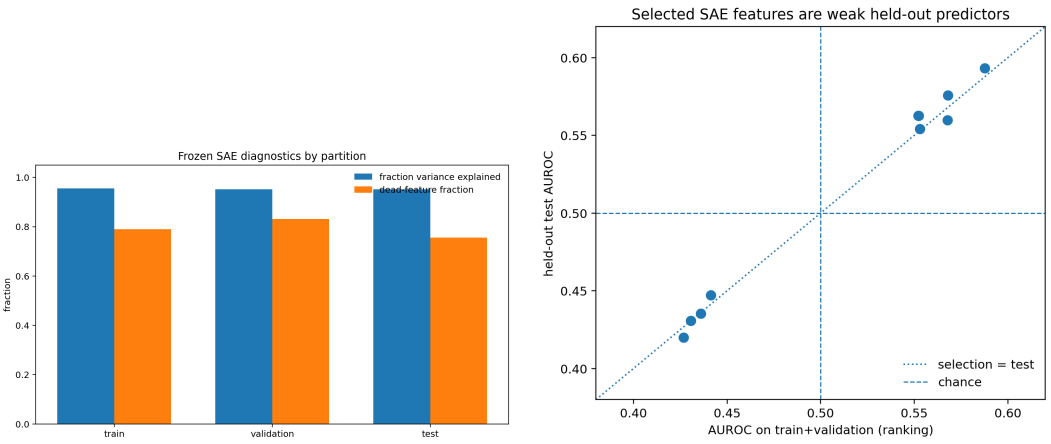

Figure 6: Sparse-autoencoder results at explicit-condition layer 30. Left: held-out reconstruction quality and dead-feature fraction by partition; mean L_0 is 64 and is reported in text. Right: held-out performance of features selected using training plus validation data. Strong reconstruction coexists with weak individual label associations.

G Reproducibility Details

Subject model. The subject model is `meta-llama/Llama-3.1-8B-Instruct`. Both model and tokenizer are pinned to revision `0e9e39f249a16976918f6564b8830bc894c89659`. We use bfloat16 weights and store extracted activations in float32.

Split and probes. The global seed is 2026. Splits are grouped by `split_group_id`, stratified by label, with 20% test and 10% validation fractions. Logistic probes use $C = 1.0$, class weighting, and a maximum of 5000 optimization iterations. Layer selection is validation-only. Grouped bootstrap summaries use 1000 resamples.

SAE. The primary SAE uses 32,768 features, TopK activation with $k = 64$, 60 epochs, batch size 1024, learning rate 3×10^{-4}, and training seed 0. Stability uses seeds 0, 1, and 2.

Interventions. The steering direction is fitted on the probe training partition. Interventions occur at the validation-selected layer with prompt-end timing. The dose grid is $\{0.5, 1, 2, 4\}$, and the held-out causal evaluation size is 500 examples. The forced-choice options are the model tokens corresponding to " Yes" and " No", with " Yes" treated as the positive option.

Execution sequence. The strict pipeline order is `validate-data` → `make-splits` → `run-baselines` → `collect-activations` → `train-probes` → `analyze-confounds` → `train-sae` → `run-interventions` → `make-figures` → `audit`. The neutral-state condition omits SAE training and live interventions by design because it is used only as an instruction-removal probe ablation and provides no independent forced-choice endpoint.

Artifact checks. Strict runs record the dataset checksum, split manifest, resolved configuration, run metadata, activation provenance, prompt hashes, model/tokenizer revisions, and figure-source checksums. Publication figures are generated from recorded result artifacts with source-data CSVs and manifest entries. Repository documentation reports both explicit and neutral strict audits passing and 254 regression tests passing; we do not treat the test-count claim as an independently rerun result in this manuscript.