EVIDENCE STATES SURVIVE WHERE ANSWERS FAIL: DECODABILITY AND CAUSAL ACCESS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A record can support a claim, its explicit negation, both, or neither. Preserving these distinctions is essential for evidence reporting. We separate evidence-state recoverability, native reporting, and causal access using entity-role swaps that preserve the exact word multiset across all four support states. Across three instruction-tuned model families, the output contrast between absent and dual support is attenuated relative to the contrast between opposing one-sided support. On held-out signed records, same-query linear decoders distinguish absent from dual support with 87.5–100% accuracy, compared with 50.0–75.0% for native yes/no readout. This advantage persists over scalar decoders trained on identical examples and state labels. Residual interventions reveal task-conditioned causal access: a within-query state intervention in Mistral exceeds an equal-norm random control by 1.79 log-odds units. Averaged over both complementary questions, transferred directions underperform random controls on signed records. On unseen source- and time-scoped records, Mistral retains strong hidden-state decoding, while Qwen and Olmo show weaker transfer. Together, these findings establish a recoverability–reporting gap and identify question- and model-dependent boundaries on its causal portability.

1 INTRODUCTION

An evidence log and a factual verdict answer different questions. If two reports support opposing claims, a complete log should retain both reports rather than resolve the conflict by omission. If neither report is present, the log should retain that absence rather than invent an opposing claim. The two cases can produce the same lack of a preferred conclusion while requiring different downstream actions: acquire evidence in one case, reconcile sources in the other. Belnap–Dunn four-valued semantics distinguishes absence, one-sided support, and inconsistent support (Dunn, 1976; Belnap, 1977); bilateral systems operationalize related distinctions (Allen et al., 2025). Whether instruction-tuned language models preserve these distinctions during computation is an empirical question.

Answer errors can originate at several points in evidence processing: entity binding, state encoding, or the mapping from a represented state to a response. Separating these possibilities requires measurements of both the answer and the computation that produces it. Work on latent truth and knowledge already demonstrates that hidden representations can predict labels more accurately than overt responses (Burns et al., 2023; Marks & Tegmark, 2024). Evidence-strength decoding also exhibits a gap between recoverable information and stated judgments (Arasteh, 2026). These observations motivate a sharper question: can we separate the evidence supplied by a record from the answer demanded by a question, and test recoverability and causal access under that separation?

We introduce a controlled evidence-state study built around two support bits, a and b. Their balance, $d = a - b$, separates opposing one-sided states. Their availability coordinate, $k = a + b - 1$, separates no support from dual support. We vary these bits by swapping entity roles while preserving the exact word multiset, including negations, within each four-state family. Each forward pass receives one independently posed question and supplies both the native answer scores and the hidden feature used by a supervised decoder. This design varies entity–predicate binding under fixed lexical counts and aligns representation and answer measurements within the same computation.

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.

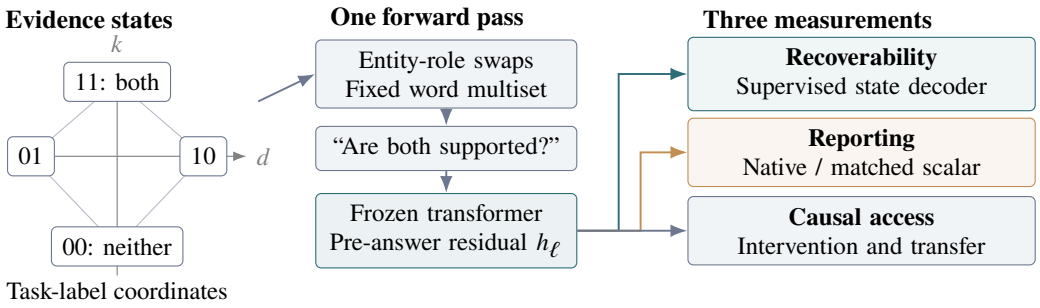


Figure 1: **Evidence, reporting, and access are separate experimental objects.** The diamond locates the four externally defined evidence states in balance–availability coordinates. Entity-role swaps change support while preserving word counts. A single question produces both the native answer distribution and the hidden activation used by a separately trained decoder. Interventions and transfer records test the functional portability of the recovered state.

Three independently developed instruction-tuned families—Qwen2.5, Mistral, and Olmo 3—show a consistent asymmetry between the two evidence contrasts. On held-out signed records, same-query decoders recover the absent-versus-dual distinction substantially more accurately than native readout. The advantage persists over scalar classifiers trained on identical examples and state labels. Evidence-state information thus remains accessible within the very computation whose answer reports it unreliably.

We test the functional role of this information through coordinate interventions and transfer. Within-query interventions shift Mistral's responses toward the donor state beyond matched random perturbations. Averaged across NEITHER and BOTH, transferred directions yield weaker signed-record effects than random controls; pooled state directions show condition-dependent performance against direct answer steering. On a separate source- and time-scoped task, Mistral retains a large hidden advantage while Qwen and Olmo transfer less effectively. These contrasts distinguish linear recoverability, causal portability, and task transfer as separate properties of evidence processing.

Our contributions are a controlled evidence-binding construction; a matched, same-query comparison of hidden, scalar, and native readouts; and counterfactual interventions that distinguish within-question access from cross-question portability. Together they support a precise conclusion: *evidence states can survive native answer failure, while their causal use remains task- and model-dependent*. Figure 1 summarizes the separation that organizes the study.

2 RELATED WORK

Evidence, binding, and knowledge conflict. Belnap–Dunn semantics treats support for a proposition and its negation independently (Dunn, 1976; Belnap, 1977). Our four labels instantiate this support interpretation; the experiment measures record binding and reporting rather than a full entailment calculus. Bilateral reasoning (Allen et al., 2025), proof generation (Tafjord et al., 2021), and conflict–ignorance decompositions (Huang et al., 2026) provide complementary perspectives. Binding studies identify mechanisms associating entities with attributes in context (Feng & Steinhardt, 2024). We preserve lexical counts while varying those associations. Retrieval-augmented generation combines external records with parametric knowledge (Lewis et al., 2020); knowledge-conflict studies examine disagreement between these sources and among contexts (Longpre et al., 2021; Xu et al., 2024). Our source task isolates which reports qualify as evidence before evaluating the resulting state.

Recoverability and response. Linear probes measure information accessible to a specified decoder (Alain & Bengio, 2016). Latent-knowledge and truth-direction studies distinguish recoverable labels from overt responses (Burns et al., 2023; Marks & Tegmark, 2024), while self-evaluation studies relate model confidence to answer correctness (Kadavath et al., 2022). A close clinical study finds evidence strength recoverable from representations but poorly reflected in stated grades, with lexical

and transfer limitations (Arasteh, 2026). We isolate externally defined four-state evidence under exact word-multiset matching, compare hidden and scalar readouts of the same query, and test causal portability across complementary questions.

Representation engineering and interventions. Latent steering vectors (Subramani et al., 2022), activation addition (Turner et al., 2023), representation engineering (Zou et al., 2023), and contrastive activation addition (Rimsky et al., 2024) manipulate behavior through activations. Inference-time interventions and layer-contrast decoding aim to improve factual responses (Li et al., 2023; Chuang et al., 2024). ROME instead locates and edits parametric factual associations through weight changes (Meng et al., 2022). We keep weights fixed and use recipient–donor residual replacement to test state access. Causal abstraction and alignment methods formalize links between variables and neural computations (Geiger et al., 2021; 2024); patching analyses emphasize intervention sites, metrics, and competing explanations (Zhang & Nanda, 2024; Makelov et al., 2023; Wu et al., 2024). Cross-query and answer-direction controls measure the functional stability of our learned coordinates.

Readout and probe controls. Probe performance depends on supervision, decoder capacity, and surface cues (Hewitt & Liang, 2019; Voita & Titov, 2020). We match state labels and training examples, retain text and random-label controls, and report native candidate scores separately. Contextual and domain calibration address response priors (Zhao et al., 2021; Fei et al., 2023); full-label calibration addresses unequal candidate lengths (Sanz-Guerrero & von der Wense, 2025). Our complete single-token candidates share one next-token distribution. Independently posed questions also separate this task from suppression across sequentially generated labels (Ma et al., 2025).

Instruction tuning and reporting behavior. Human-feedback training changes instruction following (Ouyang et al., 2022), and behavioral evaluations identify preference-sensitive responses, including sycophancy (Perez et al., 2023; Sharma et al., 2024). These studies motivate attention to reporting in instruction-tuned deployments. Our interventions manipulate inputs and activations in fixed checkpoints, addressing evidence-state access rather than the training-stage origin of a reporting preference.

3 Controlled Evidence States

3.1 States and questions

For a record T and two target statements (q_1, q_2), define

$$a = \mathbf{1}[T \text{ supports } q_1], \quad b = \mathbf{1}[T \text{ supports } q_2]. \tag{1}$$

In SIGNED, q_2 is the explicit negation of q_1. The task records support for each target independently, distinguishing an absent statement from an explicitly supported negation. In NEUTRAL, the targets have different, nonopposing predicates. QUOTED expresses the signed contents as attributed reports. Define

$$d = a - b, \qquad k = a + b - 1. \tag{2}$$

These are coordinates of the four task labels. In particular, 00 and 11 have equal balance $d = 0$ but opposite availability $k = -1, +1$.

Each call asks one of four questions: FIRST asks for a, SECOND for b, NEITHER for $\neg(a \vee b)$, and BOTH for $a \wedge b$. The SUPPORT frame asks what the record supports; LOOKUP asks whether the complete target sentence occurs. Both use the same gold labels for direct records. For the balanced endpoint task 00 versus 11, NEITHER and BOTH have opposite answer mappings. Across all four states, each has one positive and three negative states. We therefore report balanced accuracy and per-state outcomes; constant-no prediction provides the 75% raw-accuracy and 50% balanced-accuracy references.

3.2 Exact word-multiset matching

For each target branch, activation swaps a target and a distractor entity between two statements. For example,

Active: X is glossy. Y is triangular.
Inactive: Y is glossy. X is triangular.

The other branch applies the same construction to the second target. The four resulting records have identical word counts, entity counts, sentence counts, and negation counts within a family and representation. An independent parser computes gold labels from the rendered statements. Tokenization varies across surface representations; the invariant quantities are the string-level word, entity, sentence, and negation counts.

The construction distinguishes conflicting evidence from a factual verdict. A record containing both "X is glossy" and "X is not glossy" has dual support. The correct BOTH answer reports this evidence state. The QUOTED view expresses the same contents as attributed reports.

3.3 DATA POPULATIONS AND MODEL SELECTION

We use three family-disjoint populations. The controlled population contains 48 families split into 24 training, eight development, and 16 test families. Its test set combines eight familiar-layout families with new vocabulary and eight unseen tagged-layout families. The binding-transfer population contains 48 further families, with 12 in each cell of ordinary/novel predicates crossed with familiar/new formatting. It supports both within-query and cross-query interventions on the same records, enabling a paired comparison of intervention designs. The source-transfer population contains 48 separate source-scoped families reserved for transfer evaluation.

We evaluate Qwen2.5-7B-Instruct (Qwen et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Olmo-3-7B-Instruct (Team Olmo et al., 2025). Qwen and Mistral are evaluated on the controlled and binding-transfer populations. Olmo follows the controlled training/development/test split, and all three models are evaluated on the source-transfer population. Appendix A specifies the allocation and analytical role of each population.

4 MEASUREMENTS AND EXPERIMENTAL PROTOCOL

4.1 NATIVE OUTPUT AND EVIDENCE CONTRASTS

After the model-specific chat template and a fixed answer cue, one forward pass gives both candidate logits. Define

$$m(T, Q) = \ell(\texttt{yes} \mid T, Q) - \ell(\texttt{no} \mid T, Q). \tag{3}$$

Both labels are complete single-token continuations under the checkpoint's tokenizer. Native prediction is yes for $m > 10^{-4}$, no for $m < -10^{-4}$, and a tie otherwise. Throughout the paper, native readout denotes this candidate-restricted decision rule. We also report the total probability mass assigned to the two candidates. This protocol aligns activation and answer measurements at the same next-token decision point. Chain-of-thought prompting introduces generated intermediate steps and additional autoregressive computation (Wei et al., 2022; Kojima et al., 2022); the present comparison holds that computation fixed.

For each family, collect two independent margins at each corner, $z_{ab} = (m(T_{ab}, \text{FIRST}), m(T_{ab}, \text{SECOND}))$. The contrast ratio is

$$R = \frac{\|z_{11} - z_{00}\|_2}{\|z_{10} - z_{01}\|_2}. \tag{4}$$

The numerator measures absent-to-dual support response; the denominator measures the opposing one-sided response. The ratio is defined for a nonzero denominator. Table 1 reports both absolute norms, the raw contrast ratio, second-order interactions, and raw and row-normalized singular-value ratios. Together, these quantities characterize the directional sensitivity of the output margins. The contrast ratio measures readout anisotropy independently of any assumption about residual-stream dimensionality.

4.2 SAME-QUERY REPRESENTATION DECODING

We record post-block residuals at one-quarter, one-half, three-quarters, and full model depth. For the same-query analysis, the feature is the activation at the final actual input token immediately before

the requested answer. A separate record-end feature precedes the question; a two-query feature concatenates FIRST and SECOND activations. These features separate access to the question from the computational benefit of an additional query; their dimensions and forward-pass budgets are specified in Appendix C.

Fixed L_2 logistic decoders predict $00/11$, $01/10$, or all four states. Scaling is fitted only on training families; regularization is fixed at $C = 1$. Primary layers are fixed from development-only four-state accuracy: block 21 for Qwen and block 16 for Mistral and Olmo. Appendix C specifies the model-specific selection criteria.

To match supervision exactly, the principal scalar comparator uses the same 48 endpoint training records and binary state labels as the hidden decoder, replacing its feature with the native margin from that same call. We evaluate both logistic and fixed RBF classifiers. Balanced random-label controls test label association; word-count and word-pair baselines test surface cues; duplicate-query and dimension-matched two-query features separate dimensionality from information acquisition.

4.3 COUNTERFACTUAL ACCESS AND PORTABILITY

Let h and h_D denote recipient and donor activations for the same question. The donor has the opposite endpoint state, $00 \leftrightarrow 11$. Let the normalized rows of W be a state-availability probe w_k and a balance probe w_d, fitted on training data. A coordinate replacement uses

$$\delta_k = W^\top (WW^\top)^{-1} \begin{bmatrix} w_k(h_D - h) \\ 0 \end{bmatrix}, \qquad h' = h + \delta_k. \tag{5}$$

It changes the fitted availability coordinate while preserving the fitted balance coordinate. Equal-norm random controls are orthogonal to the fitted coordinate rows. Additional conditions use the balance coordinate, the complete donor vector, and a different-family donor with the same state.

Cross-query experiments apply coordinates learned for the complementary question to recipient and donor activations of the target question. The learned direction transfers across questions while both activations remain in the target-question context. We also train pooled state and answer directions on the same crossed examples: each combination of endpoint state and yes/no answer has 24 training examples. Pooled answer perturbations match the requested state-perturbation norm and point toward the donor's correct answer. This control compares state-supervised and answer-supervised steering at a matched perturbation norm.

For a donor answer $y_D \in \{0, 1\}$, define

$$\tau = (2y_D - 1)\,[m(h') - m(h)]. \tag{6}$$

We compare τ against the relevant control, average both questions and both exchange directions within each family, and compute 2,000 family-level bootstrap resamples. The resulting descriptive intervals quantify variation across families, with related questions and representations grouped within each family. Every eligible recipient contributes, including baseline errors. The known donor state defines the counterfactual target, allowing us to measure functional access under a controlled change in represented evidence.

5 EVIDENCE IS RECOVERABLE BEYOND NATIVE READOUT

5.1 EVIDENCE CONTRASTS REMAIN ASYMMETRIC ON HELD-OUT INPUTS

In the 16-family controlled test, the median within-family SIGNED/NEUTRAL ratio of R is 0.441 for Qwen and 0.553 for Mistral, with a lower signed ratio in all 16 families for both models. The corresponding raw SIGNED median R values are 0.438 and 0.534, compared with 0.906 and 0.962 in NEUTRAL. On the new 48-family binding population, the respective ratios are 0.499 and 0.394; the signed ratio is lower in 48 and 42 families. Row-normalized finite-difference analyses agree in direction, showing that the contrast persists after equalizing the two readout gains.

Olmo reproduces the controlled SUPPORT contrast on its 16 held-out families: median relative ratio 0.632, with 13 families in the same direction; raw median R is 0.610 for SIGNED and 0.993 for NEUTRAL. Its LOOKUP ratio is 0.837, with nine families in that direction. The SUPPORT–LOOKUP contrast therefore identifies task framing as a source of variation in evidence sensitivity.

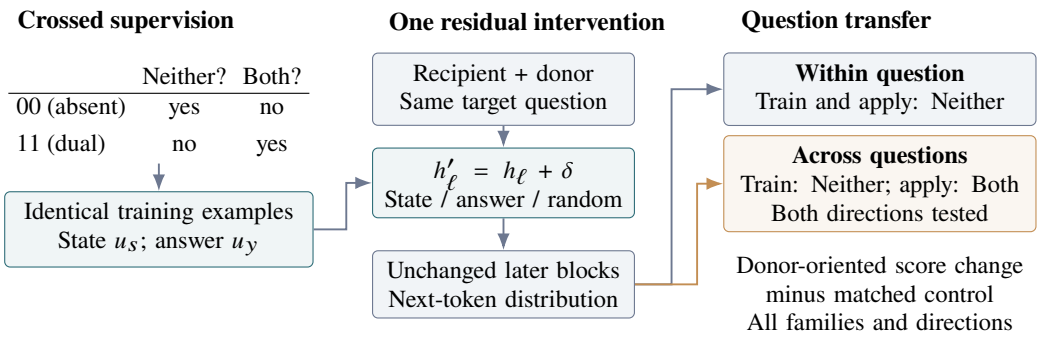


Figure 2: **Crossing evidence state with question separates state from answer supervision.** The same endpoint state requires opposite answers to NEITHER and BOTH. Interventions modify one pre-answer residual position, with unchanged downstream computation. Transferring a learned direction across questions tests a stronger property than decoding each question separately. All comparisons retain both donor directions and their matched controls.

The signed test finite-difference matrices for Qwen and Mistral retain rank two, with unequal sensitivity along the two evidence contrasts. Fixed input-independent calibration offsets cancel from Equation 4. Supervised scalar decoding provides the complementary test: how much of this anisotropic output signal supports accurate state classification?

5.2 THE GAP SURVIVES A SAME-QUERY, MATCHED-LABEL COMPARISON

Figure 3 compares the balanced $00/11$ endpoint task using a hidden feature, a matched scalar decoder, and native readout. For Qwen, the same-query hidden decoder obtains 96/96 for NEITHER and 84/96 for BOTH, compared with 61/96 and 67/96 for matched scalar logistic regression. Mistral obtains 96/96 and 94/96, whereas scalar logistic regression scores 69/96 on both questions. Native readout scores 50/96 and 55/96 for Qwen, and 72/96 and 69/96 for Mistral.

Olmo's same-query hidden classifier distinguishes the endpoint states in 32/32 held-out records for both questions. Its matched scalar classifiers score 29/32 and 27/32, while native readout scores 17/32 and 16/32. These results demonstrate strong linear decodability of absent versus dual support across held-out vocabulary and layouts. The full four-state task below measures the additional discrimination required among the one-sided states.

The fixed nonlinear scalar control also leaves a gap: its endpoint scores range from 43/96 to 71/96 for the first two models and reach 30/32 and 27/32 for Olmo. With training records and labels held fixed, the richer hidden features support more accurate state classification than the tested scalar readouts. Thus, task-relevant information remains linearly decodable from the same computation even when the native readout reports the evidence state incorrectly.

The distinction is question-dependent. On the full four-state problem, same-query hidden accuracies for Qwen are 147/192 and 142/192, and for Mistral 146/192 and 168/192, for NEITHER and BOTH respectively. Thus, robust endpoint separation coexists with task-dependent four-state discrimination. On these signed endpoints, word-count and word-pair controls score 48/96 for both Qwen and Mistral; role-normalized word pairs score 44/96. Balanced random-label classifiers score below their corresponding true-label same-query decoders (Tables 4–5). The next section examines functional access to these predictive features through interventions.

6 TASK-CONDITIONED CAUSAL ACCESS AND PORTABILITY

6.1 WITHIN-QUESTION INTERVENTIONS EXPOSE ACCESSIBLE INFORMATION

On signed records, Mistral's availability-coordinate intervention exceeds its equal-norm random control by 1.789 log-odds units, with a descriptive interval of [1.574, 2.025]. All 48 family means

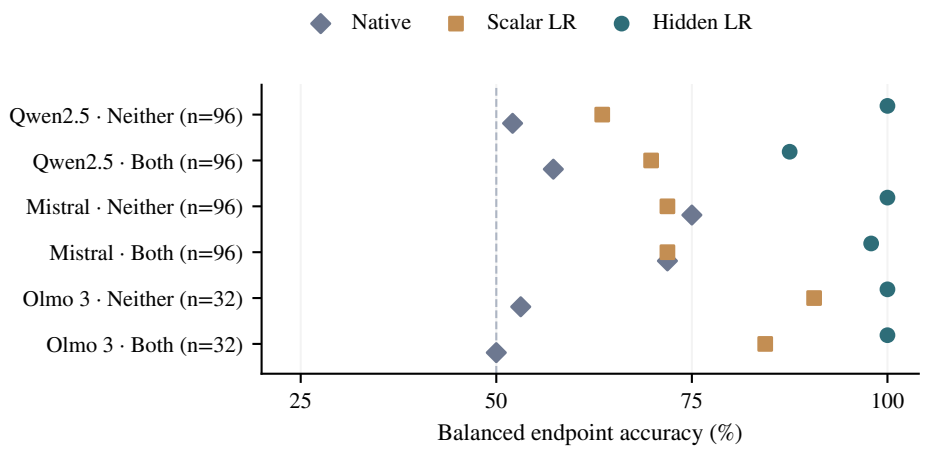


Figure 3: **Same-query hidden decoding exceeds the measured native readout on controlled signed records.** Each row evaluates the balanced $00/11$ endpoint task; hidden and scalar classifiers use identical training examples and state labels. Qwen and Mistral are tested on 48 new families ($n = 96$ endpoint records); Olmo is tested on 16 original held-out families ($n = 32$). Points show empirical accuracy in each model-specific panel; the dashed line marks the 50% balanced-task reference.

are positive, and both complementary questions have positive average effects. This establishes a controlled way to influence answers in the donor-state direction.

The interventions yield graded changes in both margins and answers. Across 192 endpoint recipients, donor-answer agreement rises from 48 before intervention to 65 after the availability patch, compared with 49 under the random control and 88 after full-vector replacement. These counts measure transfer toward the counterfactual donor; the log-odds contrast measures the accompanying continuous response. In this comparison, full-vector replacement yields greater donor-answer agreement than the availability-coordinate patch.

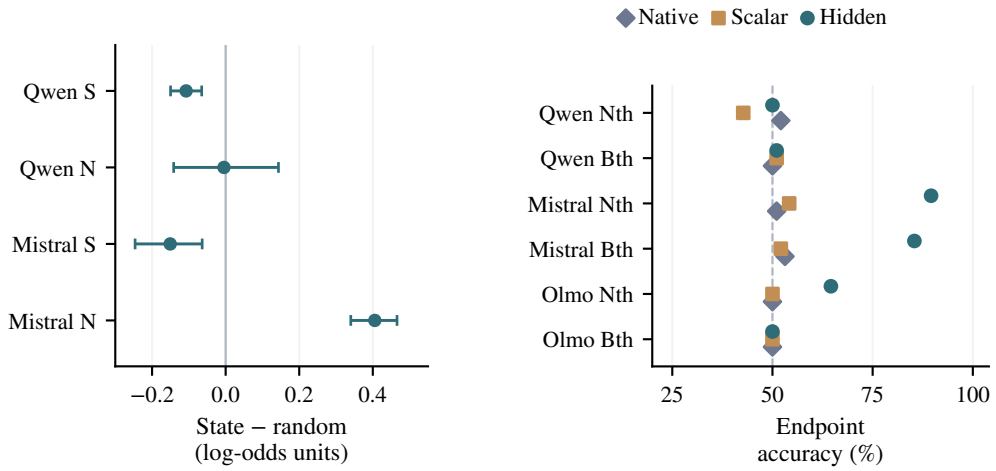
Qwen's signed average is much smaller, 0.156, with opposite question-specific tendencies. Yet the same model has a strong neutral-record availability effect of 3.422. The contrast between signed and neutral records identifies a task-dependent response within the same model. Availability-coordinate access varies jointly with target semantics, question, and model.

6.2 Question transfer delineates functional access

Figure 4a reports contrasts averaged over both complementary questions. On signed records, transferred coordinates fall below matched random controls: -0.107 for Qwen and -0.151 for Mistral. On neutral records, Mistral averages $+0.405$. Qwen's mean of -0.005 reflects opposite receiving-question effects: -1.635 for NEITHER and $+1.626$ for BOTH. For Mistral, using coordinates learned for the complementary question changes the mean signed state-minus-random contrast from positive to negative.

Pooled training provides an additional test because state and answer labels are crossed rather than fitted within one question. Averaged over both complementary questions, pooled state perturbations exceed pooled random controls in all four model–representation cells. However, comparison with an explicit equal-norm answer direction gives a different result. The state-minus-answer contrast is $+0.740$ for Mistral on neutral records, $+0.061$ on signed records with an interval spanning zero, and negative for Qwen in both representations. State-minus-answer contrasts favor state steering for Mistral on neutral records and answer steering for Qwen; the signed Mistral comparison remains unresolved at this precision.

These interventions establish task-conditioned causal influence on the answer distribution through intermediate residual states. The crossed-question and answer-direction controls delimit its functional



(a) Cross-query causal transfer.

(b) Source-task transfer.

Figure 4: **Functional portability differs across questions and tasks.** (a) Transferred state coordinates versus equal-norm random controls, averaged over both questions; bars show descriptive 95% family-bootstrap intervals. S/N denote Signed/Neutral. (b) Signed source endpoints with fixed direct-record decoders and no source fitting; Nth/Bth denote Neither/Both ($n = 96$ each). Mistral retains a hidden advantage, while Qwen and Olmo show limited transfer. Complete contrasts and source results appear in Appendices D–E.

portability: state prediction, within-question influence, and cross-question control have distinct empirical profiles. These results show that high within-question decodability need not imply cross-question controllability.

7 Transfer to Source- and Time-Scoped Records

The source population changes what makes evidence eligible. Each record contains eight attributed entries with a unit, time, and observed/planned status. Only an observed entry matching the queried unit and time supports a target. Copies of both target sentences occur outside the eligible scope in every corner, including 00. A string-presence detector can therefore find both sentences while giving the wrong evidence state. Source-scope and time-scope variation are crossed with two layouts, yielding four strata of 12 new families.

All source evaluations use fixed decoders and primary layers learned on direct records, giving a zero-additional-training transfer test. Figure 4b gives the signed endpoint results. Mistral's hidden classifier retains 86/96 accuracy for NEITHER and 82/96 for BOTH, while native readout scores 49/96 and 51/96. Matched scalar logistic regression reaches only 52/96 and 50/96. This gap is unchanged by the candidate-mass sensitivity analysis (Appendix E.1). The direct-record decoder therefore retains a substantial recoverability–reporting advantage under eligibility-based evidence reading.

Transfer varies across model families. Qwen's hidden results are 48/96 and 49/96, close to its native 50/96 and 48/96. Olmo reaches 62/96 for NEITHER and 48/96 for BOTH, compared with 48/96 native for each. Neutral records also exhibit model- and question-specific differences: Qwen's native BOTH score is 79/96 versus 76/96 for hidden decoding. The model- and question-specific pattern makes evidence-selection transfer a distinct evaluation axis.

The synthetic source records span approximately 466–580 input tokens across the three tokenizers. This suite tests structural transfer from direct statements to multi-entry scope filtering and temporal attribution. Under this combined task–interface shift (Appendix C), Mistral preserves a useful state decoder, while Qwen and Olmo exhibit greater dependence on the direct-record setting.

8 DISCUSSION

The same-query gap and the portability results motivate two complementary evaluation principles.

First, matched-supervision decoding of the same query distinguishes answer quality from the accessibility of task-relevant information. Second, cross-question transfer and direct answer-direction controls distinguish local causal influence from a portable state operation. Together, these comparisons connect a representational finding to its functional scope more sharply than either probe accuracy or a state-versus-random contrast alone.

Practical evidence readers require both state discrimination and domain transfer. A trained readout should preserve source and time constraints, distinguish missing from conflicting evidence, and support domain-specific validation and abstention.

9 CONCLUSION

Evidence states survive answer failures in a measurable, task-conditioned form. Across three instruction-tuned model families, same-query hidden features support more accurate absent-versus-dual classification than native readout under controlled entity binding. Residual interventions reveal within-question access, while cross-question and source transfer delineate its functional range. Mistral retains strong source-task decoding; Qwen and Olmo show a narrower transfer profile. Evidence recovery, reporting, and causal portability thus constitute distinct empirical capabilities.

LIMITATIONS

Model and task scope. The conclusions concern the evaluated instruction-tuned deployments: three fixed 7B checkpoints, direct evidence records, and synthetic scope filtering. Exact lexical-multiset matching controls counts while retaining entity binding. Richer syntax, multi-step deduction, longer documents, and naturally occurring source conflicts extend this task scope.

Readout and supervision. Results concern a two-token decision rule with task-dependent candidate mass (Appendix E.1). Free generation, chain-of-thought elicitation, and factual confidence concern additional decoding or calibration regimes. Hidden and scalar classifiers share training records and labels but differ in feature dimension and capacity. Their comparison measures specified readouts on finite held-out populations. Source transfer includes both new evidence-selection requirements and documented prompt-configuration variation (Appendix C).

Causal and statistical scope. Single-position residual interventions establish controlled influence; endogenous necessity, final-head localization, and training-stage origins require additional localization and training-stage comparisons. Oracle donors define counterfactual targets, whereas deployment requires an evidence estimator available at inference time. Within-query and cross-query analyses share the binding-transfer families. Their bootstrap intervals are descriptive and unadjusted. Appendix F specifies held-out, targeted, and post hoc analyses.

AI USE STATEMENT

Generative AI tools (ChatGPT and Codex) assisted hypothesis formulation, literature discovery, stimulus-generator and analysis-code development, elementary mathematical derivations, experimental planning, statistical interpretation, figure preparation, and manuscript drafting and revision. Synthetic records are instantiated by deterministic programs; an independent symbolic parser determines their support labels. Experimental measurements come from the reported model runs and saved observations. The authors take responsibility for the final text, claims, references, and experimental artifacts, including those produced with AI assistance.

ETHICS STATEMENT

The synthetic stimuli use invented entities and contain no personal data. Evidence-state reports describe eligible records; source reliability and factual verification remain separate responsibilities. Applications should preserve attribution and temporal scope, distinguish absent from conflicting evidence, and validate each readout on its intended distribution.

REPRODUCIBILITY STATEMENT

The anonymous supplement provides exact stimuli and tokenized inputs, checkpoint revisions, saved scores and predictions, and executable statistical and figure reconstruction. A companion standalone package supplies model inference, activation extraction, decoder fitting and intervention code, together with a small saved-activation regression fixture. Appendices A–D specify the scientific protocol; Appendix F and the package README distinguish saved-result reconstruction, cached-activation replay and fresh model execution.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Bradley P. Allen, Prateek Chhikara, Thomas Macaulay Ferguson, Filip Ilievski, and Paul Groth. Sound and complete neurosymbolic reasoning with LLM-grounded interpretations. In *Proceedings of the 19th International Conference on Neurosymbolic Learning and Reasoning*, volume 284 of *Proceedings of Machine Learning Research*, pp. 392–419, 2025. URL https://proceedings.mlr.press/v284/allen25a.html.

Soroosh Tayebi Arasteh. The strength of clinical evidence is recoverable from language model representations but not from their stated grades. *arXiv preprint arXiv:2606.29034*, 2026.

Nuel D. Belnap, Jr. A useful four-valued logic. In J. Michael Dunn and George Epstein (eds.), *Modern Uses of Multiple-Valued Logic*, pp. 5–37. D. Reidel, 1977. URL https://doi.org/10.1007/978-94-010-1161-7_2.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.03827.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.03883.

J. Michael Dunn. Intuitive semantics for first-degree entailments and 'coupled trees'. *Philosophical Studies*, 29(3):149–168, 1976. URL https://doi.org/10.1007/BF00373152.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14014–14031, 2023. URL https://aclanthology.org/2023.acl-long.783/.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.17191.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.02997.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Tao Huang, Rui Wang, Xiaofei Liu, Yi Qin, Li Duan, and Liping Jing. Detecting misbehaviors of large vision-language models by evidential uncertainty quantification. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.05535.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.11916.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2005.11401.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.03341.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of EMNLP*, pp. 7052–7063, 2021. URL https://aclanthology.org/2021.emnlp-main.565/.

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. Large language models do multi-label classification differently. In *Proceedings of EMNLP*, 2025. URL https://aclanthology.org/2025.emnlp-main.126/.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. *arXiv preprint arXiv:2311.17030*, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.06824.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2202.05262.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2203.02155.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, 2023. URL https://aclanthology.org/2023.findings-acl.847/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. URL https://aclanthology.org/2024.acl-long.828/.

Mario Sanz-Guerrero and Katharina von der Wense. Mitigating label length bias in large language models. *arXiv preprint arXiv:2511.14385*, 2025. URL https://arxiv.org/abs/2511.14385.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, 2022. URL https://aclanthology.org/2022.findings-acl.48/.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021. URL https://aclanthology.org/2021.findings-acl.317/.

Team Olmo et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248v1*, 2023. URL https://arxiv.org/abs/2308.10248v1.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of EMNLP*, pp. 183–196, 2020. URL https://aclanthology.org/2020.emnlp-main.14/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2201.11903.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to makelov et al. (2023)'s "interpretability illusion" arguments. *arXiv preprint arXiv:2401.12631*, 2024.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of EMNLP*, pp. 8541–8565, 2024. URL https://aclanthology.org/2024.emnlp-main.486/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16042.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

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

A STIMULI, SCOPE, AND SPLITS

A.1 A COMPLETE DIRECT-RECORD FAMILY

The following schematic family illustrates the construction. The targets are "X is glossy" and "X is not glossy." The words and their counts are identical across the four states:

00: Y is glossy. X is triangular.
Z is not glossy. X is not wooden.

10: X is glossy. Y is triangular.
Z is not glossy. X is not wooden.

01: Y is glossy. X is triangular.
X is not glossy. Z is not wooden.

11: X is glossy. Y is triangular.
X is not glossy. Z is not wooden.

Experimental families instantiate this construction with distinct names and attributes under a fixed generator. A NEUTRAL family substitutes a nonopposing second predicate, with its own matched distractor branch. A QUOTED family places the signed propositions in attributed reports. The generator varies statement order and formatting. An independent parser assigns each record its support bits from the rendered text, and exact-count checks verify the four-corner matching constraint.

Matching holds within each representation; surface forms and tokenizer segmentations vary across SIGNED, QUOTED, and NEUTRAL. Entity-role swaps make relational attribution the discriminating information.

A.2 INSTRUCTION AND QUESTION SEMANTICS

The support instruction restricts judgments to the supplied record, treats entities separately, and permits support for both signed targets. An absent statement and a supported negation have distinct truth conditions. Lookup instructions ask for complete sentence occurrence and retain the negation as part of the target text. Prompts use the checkpoint-native conversation template, including its default system-message handling. A scored assistant prefill immediately precedes the answer token. Each question receives an independent forward pass, with the configuration-specific cues described in Appendix C.

FIRST and SECOND each mention their own target rather than presenting an answer for the other target. The complement questions mention both targets because their truth conditions require both. Each is evaluated directly from the record in its own call. Their four-state truth table is:

State	First	Second	Neither	Both
00	no	no	yes	no
10	yes	no	no	no
01	no	yes	no	no
11	yes	yes	no	yes

The binary evidence-state target remains 00 versus 11 even when the corresponding yes/no answer reverses. State decoding and answer prediction are therefore evaluated with the correct mapping for each question.

A.3 FAMILY-LEVEL SPLITS AND EVALUATION DESIGN

The first population's 48 families yield 576 record bodies: 48 families times four corners times three representations. Two task frames and four questions produce 4,608 main prompts per model. Sixteen explicit-state control prompts are evaluated separately. All variants of a family share the same train/development/test membership. Eight test families use familiar formatting with new vocabulary; the other eight use a new tagged format.

The second population contains 48 new entity-role and vocabulary instances with four equally sized strata: familiar versus numbered layout crossed with ordinary versus novel predicate vocabulary. It

supplies the new-binding tests and within-query interventions for Qwen and Mistral. The within-query and cross-query interventions use this same binding-transfer population, yielding paired analyses of different intervention designs. Olmo uses the controlled-population split, with training-only fitting and development-only layer selection. Its held-out and source evaluations apply these fixed choices.

The source population is entirely separate: 48 families, four corners, and two target representations give 384 bodies. Each family is crossed with four questions and three models. Source records are reserved for evaluation of fixed direct-record predictors. Their results quantify zero-additional-training transfer to scope-aware evidence reading.

A.4 Source-record eligibility

Each source record has eight entries. Every entry states a unit, a time, a mode (observed or planned), and a quoted content sentence. The question specifies the target unit and time. An entry supplies evidence only when its scope matches and its mode is observed. Source-scope and time-scope distractors are crossed with prose-report and registry layouts.

Eligible entries satisfy all three requirements: observed status, the queried unit, and the queried time. Planned assertions and entries with mismatched units or times are out of scope. Copies of both target contents occur somewhere in every record, including a 00 record. The oracle applies all eligibility predicates before evaluating either support bit, assigning labels according to the contents of eligible reports.

This design extends direct binding to multi-entry eligibility and temporal attribution. The crossed scope and layout conditions provide a structured test of decoder transfer under new evidence-selection requirements.

B Readout Geometry and Accuracy

Let $z_{ab} \in \mathbb{R}^2$ contain the independently elicited first- and second-target margins. The four-corner finite-difference matrix is

$$J = [j_a \; j_b], \tag{7}$$

$$j_a = (z_{10} + z_{11} - z_{00} - z_{01})/2, \tag{8}$$

$$j_b = (z_{01} + z_{11} - z_{00} - z_{10})/2. \tag{9}$$

Its columns average the response to activating the first or second support bit. We retain singular values of J, the smaller-to-larger ratio, and the corresponding ratio after normalizing each nonzero row. The row-normalized version removes a fixed positive gain difference between the two readouts. The interaction term

$$z_{11} + z_{00} - z_{10} - z_{01} \tag{10}$$

is reported alongside the first-order contrasts to characterize nonadditive responses.

Subtracting a fixed vector c from every corner leaves these differences unchanged, so the contrasts are invariant to constant calibration offsets. Their magnitudes characterize response sensitivity. The scalar classifiers and two-margin controls separately measure state discrimination under supervised transformations, complementing the geometric analysis.

Ties ($|m| \leq 10^{-4}$) remain in the evaluation denominator and define an accuracy interval: the lower endpoint counts each tie as wrong and the upper endpoint counts it as correct. For imbalanced four-state questions, balanced accuracy is the average of positive-state sensitivity and negative-state specificity. On the $00/11$ endpoint tasks, class counts are equal, so ordinary accuracy is balanced accuracy. Decoder evaluation includes all assigned test records, spanning correct and incorrect native responses.

B.1 Numerical geometry across evaluation populations

Table 1 gives the quantities underlying the directional-response claims for every frame and representation in the controlled and binding-transfer tests. In the controlled SUPPORT/SIGNED test, the absent-to-dual norm has medians 15.523, 24.719, and 10.807 for Qwen, Mistral, and Olmo; the opposing-one-sided norm has medians 35.693, 45.054, and 19.450. The corresponding per-family

median ratios are all below one. The raw ratios and the within-family Signed/Neutral ratios are separate statistics.

Table 1: Absolute output contrasts and complete geometry summaries on held-out records. Each entry is a median of per-family quantities; the median ratio is not computed by dividing the two median norms. Here $u = z_{11} - z_{00}$ and $v = z_{10} - z_{01}$.

Model	Set	Frame	View	$\|u\|$	$\|v\|$	R	ρ	$\bar{\rho}$	$\|w\|$
Qwen	B48	Look.	N	37.203	37.586	0.975	0.685	0.801	9.284
Qwen	B48	Look.	Q	26.457	24.896	1.101	0.352	0.720	8.615
Qwen	B48	Look.	S	24.306	30.674	0.874	0.516	0.706	10.717
Qwen	B48	Sup.	N	41.833	40.356	1.003	0.773	0.892	7.599
Qwen	B48	Sup.	Q	9.917	19.018	0.541	0.241	0.293	8.593
Qwen	B48	Sup.	S	14.369	33.274	0.476	0.198	0.223	12.841
Mistral	B48	Look.	N	54.667	53.129	1.000	0.817	0.926	3.841
Mistral	B48	Look.	Q	39.831	42.545	0.888	0.514	0.799	19.595
Mistral	B48	Look.	S	30.569	42.650	0.738	0.434	0.596	27.308
Mistral	B48	Sup.	N	46.951	45.521	1.002	0.747	0.881	6.891
Mistral	B48	Sup.	Q	21.094	37.282	0.601	0.194	0.434	15.773
Mistral	B48	Sup.	S	18.586	41.212	0.467	0.175	0.286	24.095
Qwen	C16	Look.	N	36.423	39.560	0.943	0.581	0.771	14.095
Qwen	C16	Look.	Q	27.211	30.882	0.912	0.535	0.693	11.909
Qwen	C16	Look.	S	23.587	31.968	0.682	0.442	0.633	11.294
Qwen	C16	Sup.	N	44.515	49.340	0.906	0.814	0.864	6.772
Qwen	C16	Sup.	Q	11.159	20.972	0.602	0.180	0.239	8.829
Qwen	C16	Sup.	S	15.523	35.693	0.438	0.184	0.198	16.197
Mistral	C16	Look.	N	55.227	56.547	0.981	0.920	0.939	2.831
Mistral	C16	Look.	Q	37.816	47.946	0.852	0.539	0.699	23.487
Mistral	C16	Look.	S	23.302	43.111	0.566	0.393	0.709	27.888
Mistral	C16	Sup.	N	46.469	54.062	0.962	0.812	0.879	7.004
Mistral	C16	Sup.	Q	23.697	40.590	0.544	0.148	0.217	18.344
Mistral	C16	Sup.	S	24.719	45.054	0.534	0.165	0.208	22.257
Olmo	C16	Look.	N	17.915	18.123	0.993	0.947	0.967	1.458
Olmo	C16	Look.	Q	8.179	11.285	0.894	0.478	0.734	7.146
Olmo	C16	Look.	S	9.667	11.495	0.842	0.254	0.630	4.751
Olmo	C16	Sup.	N	27.214	26.477	0.993	0.936	0.957	1.403
Olmo	C16	Sup.	Q	8.621	13.019	0.692	0.320	0.670	7.280
Olmo	C16	Sup.	S	10.807	19.450	0.610	0.126	0.261	7.086

C16 denotes the 16-family controlled test, B48 the 48-family binding-transfer test. N/Q/S: Neutral/Quoted/Signed. ρ is the singular-value ratio of J, $\bar{\rho}$ its row-normalized analogue, and $w = z_{11} + z_{00} - z_{10} - z_{01}$. Norms use the logit-margin scale. All family-level values are supplied in `analysis/geometry_by_family.csv`.

Model	Question	n	Hidden LR	Scalar LR	Scalar RBF	Native
Qwen2.5	Neither	96	96	61	43	50
Qwen2.5	Both	96	84	67	64	55
Mistral	Neither	96	96	69	63	72
Mistral	Both	96	94	69	71	69
Olmo 3	Neither	32	32	29	30	17
Olmo 3	Both	32	32	27	27	16

Table 2: Exact correct counts on controlled signed $00/11$ endpoint records. Scalar and hidden classifiers receive identical training examples and state labels. Qwen and Mistral use the 48-family binding-transfer test; Olmo uses the 16-family controlled held-out test.

Frame	Targets	Question	Hidden LR	Scalar LR	Scalar RBF	Native
Lookup	Neutral	Neither	32	30	30	19
Lookup	Neutral	Both	32	32	32	26
Lookup	Quoted	Neither	32	29	29	21
Lookup	Quoted	Both	32	31	31	16
Lookup	Signed	Neither	32	29	29	27
Lookup	Signed	Both	32	26	22	16
Support	Neutral	Neither	32	32	32	18
Support	Neutral	Both	32	32	32	32
Support	Quoted	Neither	32	27	27	16
Support	Quoted	Both	32	29	29	16
Support	Signed	Neither	32	29	30	17
Support	Signed	Both	32	27	27	16

Table 3: Olmo held-out endpoint results ($n = 32$ per row), covering every task frame and representation.

C Models, Features, and Decoders

C.1 Model computation and activation sites

All target models are fixed instruction-tuned checkpoints evaluated without gradient updates. Qwen has 28 transformer blocks, and Mistral and Olmo each have 32 in the evaluated configurations. The four sampled depths are therefore 7/14/21/28 for Qwen and 8/16/24/32 for the others. The recorded features are the residual activations immediately following each sampled transformer block.

The same-query activation is taken at the final nonpadding input position immediately before the next-token answer. Record-end states are taken before the question text, with a single shared-prefix feature representing each record and frame. Native and probe predictions from one question share the same model forward. All evaluations use the checkpoint-native chat template and a scored assistant prefill. The controlled Qwen/Mistral configuration additionally includes a user-side answer cue; the controlled Olmo and source-task configurations use the assistant prefill alone. Each configuration is fixed within its panel, and every same-query comparison shares the full input and prefill. Source transfer therefore spans both a record-task change and this documented interface variation. A two-query concatenation explicitly consumes two question-conditioned computations, whether or not a later projection compresses its feature dimension.

Inference uses evaluation mode, eager attention, nonquantized BF16 weights and a batch size of one. Padding and causal positions are fixed within the required comparisons. Both candidate logits are read from the same next-token distribution. Repeated identical inputs and identity interventions produce matching score and sampled-activation measurements in the tested comparisons.

C.2 Supervised estimators and feature budgets

Every scaler is fitted on training records only. Logistic classifiers use L_2 regularization with $C = 1$, the L-BFGS solver, tolerance 10^{-4}, and at most 1,000 iterations. The scalar nonlinear comparator is an RBF SVM with $C = 1$ and the implementation's scale-based kernel width, fixed before evaluation. Both fixed estimators are reported for every evaluation condition.

Binary availability decoders train on 48 examples: the two endpoint corners from each of 24 training families. Binary balance decoders use the two one-sided corners. Four-state decoders use 96 examples. For each endpoint comparison, hidden and scalar estimators use identical training records, state labels, and evaluation records.

Feature controls distinguish information acquisition from dimension. Single-query hidden vectors consume one forward; concatenated first/second vectors consume two. Compressed concatenations retain the two-call budget at a single-vector dimension; duplicated single-query features provide the complementary one-call, double-dimension control. Text baselines include word counts and word pairs, including target-role-normalized variants. Unlike word counts, word-pair features can capture entity–predicate binding patterns.

The balanced binary random-label control assigns original and reversed mappings to exactly 12 families each. Its four-state counterpart assigns each cyclic mapping to six families. This construction balances the association between control labels and true states across training families. All task and family variants remain grouped during evaluation.

C.3 LAYER SELECTION AND TEST USE

Qwen and Mistral retain blocks 21 and 16 from the initial development selection on concatenated FIRST/SECOND pre-answer features. The selection criterion averages four-state accuracy over six frame–representation cells. These blocks remain fixed for the single-query, intervention, and transfer analyses. Olmo selects a single primary block by mean development four-state accuracy over the 24 frame/representation/question cells. Development means are 0.306, 0.855, 0.844, and 0.839 at blocks 8, 16, 24, and 32 respectively. The chosen block 16 and all estimators are fixed before the 16 test families or any source family is evaluated. Table 6 reports the test scores at all four sampled blocks for SUPPORT/SIGNED. The complete frame, representation, question, block and stratum results are in `analysis/decoder_results_expanded.csv`.

C.4 OBSERVED CONTROL AND LAYER RESULTS

Table 4 places the balanced random-label controls beside the corresponding true-label decoders. On the key NEITHER/BOTH endpoints, random-label accuracies are 57/96 and 68/96 for Qwen, 55/96 and 45/96 for Mistral, and 22/32 and 21/32 for Olmo. Table 4 reports the measured accuracies of the fixed label-permutation controls.

Table 5 separates query acquisition, feature dimension, and text-only prediction. In the Qwen/Mistral signed binding test, concatenated and compressed two-query features each reach 96/96, as do duplicated-First features. These acquisition controls complement the same-query results: the NEITHER advantage in both models is already present in a single forward. Fixed output-feature decoders and all text baselines are shown together with their input dimensions and call budgets.

Table 6 makes the sampled-layer comparison explicit. Qwen's BOTH endpoint score rises from 84/96 at the primary block to 91/96 at the last block; the primary-block comparisons throughout the paper retain the development selection. Olmo's endpoint scores reach 32/32 at each of the three deeper sampled blocks, while its four-state scores remain question-dependent. The complete data retain every sampled block and all task conditions.

Table 4: Same-query evidence-state decoding and balanced random-label controls. All rows evaluate the balanced absent-versus-dual endpoint task under Support/Signed; control classifiers are scored against the true held-out state labels.

Model	Question	Block	True-label LR	Random-label LR
Qwen	First	21	96/96	58/96
Qwen	Second	21	96/96	61/96
Qwen	Neither	21	96/96	57/96
Qwen	Both	21	84/96	68/96
Mistral	First	16	95/96	54/96
Mistral	Second	16	96/96	49/96
Mistral	Neither	16	96/96	55/96
Mistral	Both	16	94/96	45/96
Olmo	First	16	32/32	21/32
Olmo	Second	16	32/32	20/32
Olmo	Neither	16	32/32	22/32
Olmo	Both	16	32/32	21/32

Table 5: Information-acquisition, feature-dimension and shallow-text controls on the Support/Signed binding-transfer endpoints. Each score uses the same 96 evaluation records; training uses 48 endpoint records. Single-position and concatenated features retain their distinct acquisition budgets.

Feature	Fit	Calls	Dimensions (Q/M)	Qwen	Mistral
Record end	LR	1	3584/4096	85/96	74/96
First only	LR	1	3584/4096	96/96	95/96
Second only	LR	1	3584/4096	96/96	96/96
Two-query concatenation	LR	2	7168/8192	96/96	96/96
Two-query compressed	LR	2	3584/4096	96/96	96/96
Duplicated First	LR	1	7168/8192	96/96	96/96
Two output margins	LR	2	2/2	74/96	73/96
Two output margins	RBF	2	2/2	81/96	72/96
Four output margins	LR	4	4/4	83/96	84/96
Four output margins	RBF	4	4/4	86/96	83/96
Word counts	LR	0	vocabulary	48/96	48/96
Word pairs	LR	0	vocabulary	48/96	48/96
Role-normalized word pairs	LR	0	vocabulary	44/96	44/96

Calls count language-model forward passes used to obtain the feature, not CPU classifier evaluation. Dimensions distinguish Qwen/Mistral. Word-pair representations use the record text; target-role normalization also receives target identity. These text controls are available for Qwen/Mistral panels; Table 4 gives Olmo's measured random-label control.

Table 6: Test accuracy at every sampled block under Support/Signed. The first four score columns classify the balanced 00/11 endpoints using one question's activation; the final two classify all four states. Asterisks mark development-selected primary blocks.

Model	Block	First	Second	Neither	Both	Neither (4)	Both (4)
Qwen	7	57/96	45/96	50/96	49/96	47/192	49/192
Qwen	14	54/96	56/96	52/96	57/96	59/192	57/192
Qwen	21*	96/96	96/96	96/96	84/96	147/192	142/192
Qwen	28	96/96	95/96	94/96	91/96	146/192	155/192
Mistral	8	49/96	45/96	50/96	52/96	52/192	51/192
Mistral	16*	95/96	96/96	96/96	94/96	146/192	168/192
Mistral	24	93/96	96/96	96/96	92/96	168/192	165/192
Mistral	32	92/96	96/96	95/96	91/96	155/192	150/192
Olmo	8	20/32	20/32	18/32	21/32	16/64	18/64
Olmo	16*	32/32	32/32	32/32	32/32	56/64	56/64
Olmo	24	32/32	32/32	32/32	32/32	57/64	54/64
Olmo	32	32/32	32/32	32/32	32/32	57/64	56/64

Qwen/Mistral use the binding-transfer test; Olmo uses its 16-family controlled test. Every frame, view, question, sampled block and test stratum is tabulated in `analysis/decoder_results_expanded.csv`; none is selected using these test scores.

D Intervention Details and Complete Contrasts

D.1 Directions in raw hidden coordinates

Probe coefficients are converted back through the training scaler before normalization. For the full-row-rank probe matrix W in Equation 5, the Gram matrix $WW^\top$ defines the minimum-norm change that realizes the requested availability displacement while preserving the fitted balance coordinate. The geometry is computed in floating-point host arithmetic; the effective BF16 change at the intervention site is separately recorded. The two probe rows define the intervention coordinates, allowing one fitted state component to change while the other remains fixed.

Random controls are generated deterministically for each recipient and projected into the relevant fitted nullspace, then normalized to the requested perturbation norm. Availability and balance

perturbations each have their own equal-norm random comparator, supporting separate within-coordinate effect estimates. Full-vector replacement and same-state, different-family donors provide full-state and background references.

D.2 Cross-question and pooled controls

For cross-question replacement, the source query is always the complement: NEITHER supplies coordinates to BOTH, and vice versa. The transferred coordinate, sign, layer, and replacement rule are fixed. Both h_D and h belong to the target question. The same-state donor is a fixed different family within the same stratum.

Pooled state and pooled answer classifiers receive the same 96 training vectors: 24 families, two endpoint states, and two complementary questions. All four state-by-answer cells contain 24 examples. Let u_s, u_y be their normalized raw-coordinate coefficients. The state replacement is

$$\delta_s = u_s u_s^\top (h_D - h). \tag{11}$$

The answer comparator uses

$$\delta_y = \|\delta_s\|_2 (2y_D - 1) u_y. \tag{12}$$

The donor's answer determines the sign of the answer-steering perturbation, providing a norm-matched comparator with direct response supervision. The pooled random comparator lies in the nullspace of u_s and u_y at the same norm. State-minus-answer and state-minus-random contrasts therefore measure distinct aspects of the intervention's specificity.

D.3 Family aggregation

The primary cross-query mean uses four endpoint cases per family: two questions times two opposite donor directions. A condition contrast is formed within each case and then averaged within the family. The bootstrap samples 48 families with replacement, keeping all related conditions together; 2,000 shared resamples are used for the reported contrasts. The intervals are descriptive and unadjusted for multiple comparisons. Table 7 reports the complete model–representation set. Receiving-question means are also supplied in `analysis/causal_by_question.csv`; each averages the two donor directions within a family and then the 48 families. These descriptive decompositions retain the original question-averaged estimand.

Model	Targets	Contrast	Mean	Descriptive 95% interval	Positive families
Qwen2.5	Neutral	Pooled state − answer	-4.043	[-4.418, -3.664]	0/48
Qwen2.5	Neutral	Pooled state − random	+1.488	[+1.281, +1.697]	48/48
Qwen2.5	Neutral	Cross-query state − random	-0.005	[-0.141, +0.144]	23/48
Qwen2.5	Signed	Pooled state − answer	-0.186	[-0.226, -0.147]	2/48
Qwen2.5	Signed	Pooled state − random	+0.065	[+0.005, +0.124]	32/48
Qwen2.5	Signed	Cross-query state − random	-0.107	[-0.150, -0.065]	7/48
Mistral	Neutral	Pooled state − answer	+0.740	[+0.662, +0.823]	48/48
Mistral	Neutral	Pooled state − random	+1.128	[+1.022, +1.233]	48/48
Mistral	Neutral	Cross-query state − random	+0.405	[+0.340, +0.466]	46/48
Mistral	Signed	Pooled state − answer	+0.061	[-0.013, +0.128]	27/48
Mistral	Signed	Pooled state − random	+0.697	[+0.567, +0.822]	45/48
Mistral	Signed	Cross-query state − random	-0.151	[-0.246, -0.064]	12/48

Table 7: All primary cross-query and pooled-control contrasts in log-odds units. Each family contributes an average over both questions and both opposite-endpoint exchanges before resampling. Positive-family counts indicate the direction of the mean contrast within each family.

E Source Transfer and Candidate Mass

Table 8 includes signed and neutral source endpoints, both questions, and both fixed scalar models. The hidden estimator is the fixed same-query availability decoder trained on direct records. All source predictions use the direct-task parameters and selected layer, measuring transfer without additional training or calibration.

Mistral's signed hidden scores of 86/96 and 82/96 retain a substantial endpoint distinction relative to scalar and native readouts. Qwen and Olmo show smaller transfer gains on signed records. The neutral results also vary by question and model, including conditions where native readout outperforms the fixed hidden classifier. Table 8 presents these complementary outcomes together.

Model	Targets	Question	Hidden LR	Scalar LR	Scalar RBF	Native
Qwen2.5	Neutral	Neither	48	38	26	69
Qwen2.5	Neutral	Both	76	74	54	79
Qwen2.5	Signed	Neither	48	41	56	50
Qwen2.5	Signed	Both	49	49	49	48
Mistral	Neutral	Neither	69	64	65	49
Mistral	Neutral	Both	78	49	58	49
Mistral	Signed	Neither	86	52	47	49
Mistral	Signed	Both	82	50	50	51
Olmo 3	Neutral	Neither	62	54	54	48
Olmo 3	Neutral	Both	48	49	49	48
Olmo 3	Signed	Neither	62	48	48	48
Olmo 3	Signed	Both	48	48	48	48

Table 8: All source-task endpoint comparisons under the support frame. Every cell has $n = 96$ balanced endpoint records from 48 new families. All predictors retain their direct-record fits and selected layers.

E.1 Candidate probability diagnostics

The yes/no margin uses logits from one conditional distribution. We retain each token's probability and their total mass. Table 9 reports prompt lengths and candidate mass for the third-model controlled panel and all source panels. Mistral assigns total mass below 0.1 in 284 of 1,536 source prompts: 149 belong to FIRST and 135 to SECOND. All 384 NEITHER and all 384 BOTH prompts exceed 0.1.

High-mass sensitivity analysis. Using a fixed threshold $M = p(\texttt{yes}) + p(\texttt{no}) > 0.1$, we recompute the source endpoint comparisons from the saved per-record predictions. All 96 records in each model–representation–question cell of Table 8 are retained. In particular, Mistral's signed hidden scores remain 86/96 and 82/96, versus native scores of 49/96 and 51/96. Its minimum endpoint mass is 0.3194 for NEITHER and 0.1636 for BOTH, across target representations. Thus, the source endpoint decoding advantage is present entirely in the high-mass sample. Full-sample evaluations remain primary; the threshold analysis is a post hoc sensitivity check.

Conditioning the two probabilities on their sum gives log odds $\log[(p_y/M)/(p_n/M)] = \log(p_y/p_n) = m$. This identity preserves the sign under pairwise normalization, but measures a restricted candidate decision rather than the content of unrestricted generation. The low-mass FIRST/SECOND cases remain in the four-corner geometry and are reported as such.

Panel	Model	Min. tokens	Median tokens	Max. tokens	Median mass	Mass < 0.1
Controlled	Olmo 3	185	212.0	284	0.9966	0/4608
Source	Qwen2.5	485	506.0	537	0.9996	0/1536
Source	Mistral	519	550.5	580	0.3773	284/1536
Source	Olmo 3	466	487.0	518	0.9996	0/1536

Table 9: Prompt lengths and total probability on the two answer candidates, over all main questions and states in the corresponding panels. All observations, including low-mass cases, contribute to the reported results.

F Reproducibility and Analysis Design

Reconstruction from saved outputs. The supplement provides exact scored inputs, per-record predictions, intervention effects, and resampling indices for reconstruction of the reported results.

The reconstruction scripts reproduce the geometry, decoder comparisons, intervention contrasts, and candidate-mass analyses. The README maps each result to its inputs and script, including the complete tables underlying Tables 1–6.

Independent model inference. The companion `anonymous_reproduction` package provides a standalone command-line interface for pinned public-checkpoint loading, exact prompt construction, single-forward activation extraction, train-only decoder fitting, development-set layer selection and residual interventions. Its README specifies the complete execution order and separate inference environments. A fresh run creates its own activation cache from the public weights; original activation archives are optional. Checkpoint revisions, dependencies, numerical checks and selection seals are included.

Cached-activation analysis and verification. The standalone package includes a saved-activation regression fixture for replaying predictions and requested and effective interventions. The `evaluate` and `audit` commands also operate on complete caches generated by fresh inference. Public weights are obtained separately; historical activation caches are optional. Validation covers input reconstruction, saved-output analysis, and the supplied activation fixture; the exported end-to-end GPU pipeline has not been rerun.

The study separates held-out model evaluation, transfer evaluation on new families, and targeted intervention comparisons on a shared binding population. The matched-sample scalar comparison for Qwen and Mistral is a post hoc control; its predictor fitting uses only the designated training families. Source-task evaluation applies fixed predictors. All related questions, representations, and interventions remain grouped by family in the statistical analysis.