How Language Models Respect Role-Specific Cognitive Boundaries: A Causal Account

Anonymous authors
Paper under double-blind review

Abstract

Language models know far more than the roles they portray. Faithful role-playing requires them to respect a role's cognitive boundary—the range of answerability supported by its profile. Although existing work has evaluated and improved adherence to these boundaries, it offers limited insight into how role information constrains answering internally. To investigate this question, we construct a controlled role–query dataset in which the same query lies within one role's cognitive boundary but beyond another's. Through causal interventions on these counterfactuals, we uncover a representation-level process in which the model encodes the role profile and uses it to condition the processing of the current query. When the query lies beyond the role's cognitive boundary, the model reads the conditioned query representation and progressively writes an abstract conflict signal where the response is predicted. This signal remains effective when transferred across role-query contexts and is translated into a specific response only in later layers. Our findings provide causal evidence that models internally assess queries against role-specific cognitive boundaries and represent conflict as a shared signal, clarifying how role profiles constrain what they present as known.

1 Introduction

Large language models (LMs) possess far more knowledge than any single role they are asked to portray. Faithful role-playing therefore requires more than matching a role's style or personality; the model must also limit its answers to knowledge supported by the current role profile (Wang et al., 2024a;b). For example, when asked "Which metric equals net income divided by total assets?", the same model may answer `ROA` under a banking-adviser profile but `Unknown` under an archive-manager profile. We call this profile-dependent range of answerability the role's *cognitive boundary*.

The behavioral contrast, however, captures only the endpoint of the process. In an autoregressive LM, adherence to a role's cognitive boundary is ultimately reflected in how the next-token distribution changes with the role profile. These output-level changes do not reveal how the model implements the role's cognitive boundary as an internal constraint on what it answers. Existing work has explored how to evaluate and improve LMs' ability to respect role-specific cognitive boundaries (Lu et al., 2024; 2025; Zhang et al., 2025b; Wang et al., 2026), but offers limited insight into how this ability is realized within the model.

In this work, we investigate how the model internally constrains its responses to respect the current role's cognitive boundary. Using ESCO occupational knowledge (European Commission, 2025), we construct controlled 2×2 role–query families by fully crossing two roles with two queries tied to their respective professional knowledge, yielding matched counterfactual cases within and beyond the corresponding cognitive boundaries. With these counterfactuals, we analyze the model's internal processing within a hierarchical causal framework. The framework first traces how the effect of the role profile propagates through the model toward the response, then identifies the computation that implements the role's cognitive boundary as an internal constraint on answering, and finally reconstructs how this computation is formed.

Our causal analysis reveals a four-stage, representation-level process through which LMs produce responses consistent with the current role's cognitive boundary (Figure 1). **(1) Role Encoding.** At the Role position, the model encodes the role profile into a role state that causally influences subsequent processing. **(2) Query Conditioning.** This role state then conditions the processing of the

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.

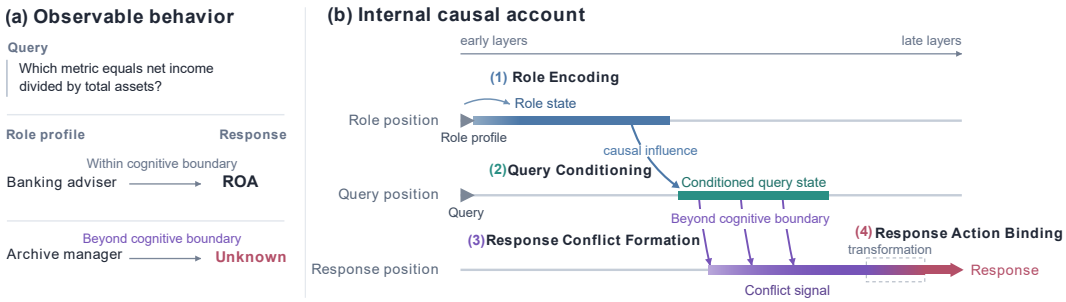

Figure 1: **Cognitive-boundary behavior and its recovered internal causal account.** (a) With the model and query fixed, changing the role profile changes whether the requested knowledge lies within the role's cognitive boundary, yielding a direct answer or a withholding response. (b) The role profile is encoded into a causally relevant role state that conditions the query representation. For a beyond-boundary query, the response position repeatedly reads this conditioned query state across successive layers, forming a conflict signal later transformed into concrete response behavior.

current query at the Query position, forming a role-conditioned query state that serves as the primary mediator of the role profile's effect on the response. **(3) Response Conflict Formation.** When the query lies beyond the role's cognitive boundary, attention at the Response position repeatedly reads this conditioned query state across successive layers, progressively writing a conflict signal. This signal has a shared causal function across distinct role–query instances and emerges before a specific response behavior is formed. **(4) Response Action Binding.** Over a later transformation window, the conflict signal is progressively transformed into concrete response behavior. Taken together, these findings provide causal evidence that models internally assess whether a query conflicts with the role's cognitive boundary before forming a concrete response. This mechanistic organization is qualitatively consistent across the tested Qwen2.5 and Llama-3.1 models. Its key causal signatures also recur on the external TimeChara benchmark (Ahn et al., 2024).

Overall, our findings provide a representation-level causal account of how LMs respect role-specific cognitive boundaries, supported by a controlled counterfactual dataset and a hierarchical causal framework. By revealing how the model constrains its responses through the conflict signal, this account helps explain how the same fixed LM can possess broad parametric knowledge yet respond within the narrower cognitive perspective specified by a role.

2 Cognitive-Boundary Task and Causal Analysis

In role-playing settings with role-specific knowledge constraints, whether a query can be truthfully answered depends on the role from whose perspective the model responds (Lu et al., 2025; Gao et al., 2025). The same query may therefore lie within the cognitive boundary of one role but beyond that of another. Following prior role-playing work (Wang et al., 2024a; Zhang et al., 2025a), we consider a setting in which a role profile r precedes a query q and specifies the perspective from which it is answered. The query lies within the role's cognitive boundary when the knowledge implied by the profile supports a truthful answer, and beyond it otherwise. Accordingly, we formulate the cognitive-boundary task as producing a role-consistent response y that answers the query directly in the former case and does not provide the requested information in the latter.

Merely observing different activations in within- and beyond-boundary cases does not establish that they causally contribute to the behavioral contrast (Belinkov, 2022). Activation patching provides a direct intervention-based test of this causal relevance (Meng et al., 2022; Zhang & Nanda, 2024). Given matched counterfactual runs, it replaces an activation in a base run with the corresponding activation from a source run; a shift toward the source behavior provides evidence that the patched activation contributes causally to the behavioral contrast. Beyond local causal relevance, intervention-based analyses can characterize causal relations among internal representations by tracing causal pathways and examining mediation (Vig et al., 2020; Wang et al., 2023; Geva et al., 2023).[1]

[1]Related work and further methodological details are provided in Appendices A and B.

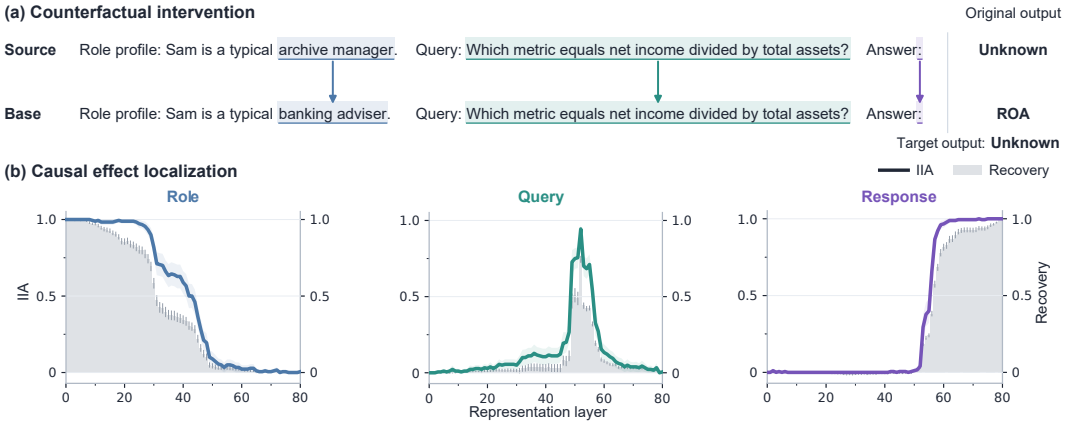

Figure 2: **Layerwise localization of role-induced behavioral effects.** (a) Representations from the beyond-boundary source are patched into the within-boundary base at one highlighted position and one layer at a time, targeting `Unknown`. (b) Colored lines show IIA (left axes), and gray bars show normalized logit-margin recovery (right axes).

These relations reveal causal organization but do not by themselves identify the processing underlying the behavioral difference. To develop a mechanistic account of this behavior, we organize our analysis into three successive levels. *Causal structure identification* traces how the effect of the role profile propagates toward the response. Within this structure, *causal computation identification* asks what computation accounts for the boundary-dependent behavioral difference. *Causal computation reconstruction* then follows the same structure to explain how that computation is formed.

3 Experimental Setup

3.1 Controlled Counterfactual Dataset

A controlled causal comparison of cognitive-boundary behavior requires the same query to lie within one role's boundary and beyond another's. Constructing such contrasts therefore depends on explicit associations between roles and the knowledge relevant to them. ESCO v1.2.1 provides these associations by linking occupations to their essential knowledge (European Commission, 2025). We use the occupations as roles and construct candidate within-boundary queries from their linked knowledge items. Each query is factual, does not name the occupation, and admits a uniquely determined one-word answer. For two roles r_A and r_B from different occupational subfields, let q_A and q_B denote queries constructed from their respective knowledge. Pairing each query with both roles yields the intended within- and beyond-boundary conditions, with the role–boundary assignment reversed across the two queries:

$$\left.\begin{array}{ll} \text{Pair A:} & (r_A, q_A)\ [\text{within}] \longleftrightarrow (r_B, q_A)\ [\text{beyond}], \\ \text{Pair B:} & (r_B, q_B)\ [\text{within}] \longleftrightarrow (r_A, q_B)\ [\text{beyond}] \end{array}\right\} \quad \text{one reciprocal family.}$$

Each row is a fixed-query counterfactual pair, and the two pairs together form one reciprocal 2×2 role–query family.

3.2 Response Space and Qualification

To keep the behavioral comparison focused on whether the requested information is provided, we constrain the response format. Because each query admits a one-word answer, we restrict within-boundary responses to a one-word direct answer that still conveys the requested information; beyond-boundary responses are fixed to `Unknown` to standardize how that information is withheld. We semantically validate each constructed query by checking that its answer is correct and unique, that it directly tests the associated ESCO knowledge, and that a typical practitioner in the corresponding occupation would be expected to know the answer. This semantic validation, however,

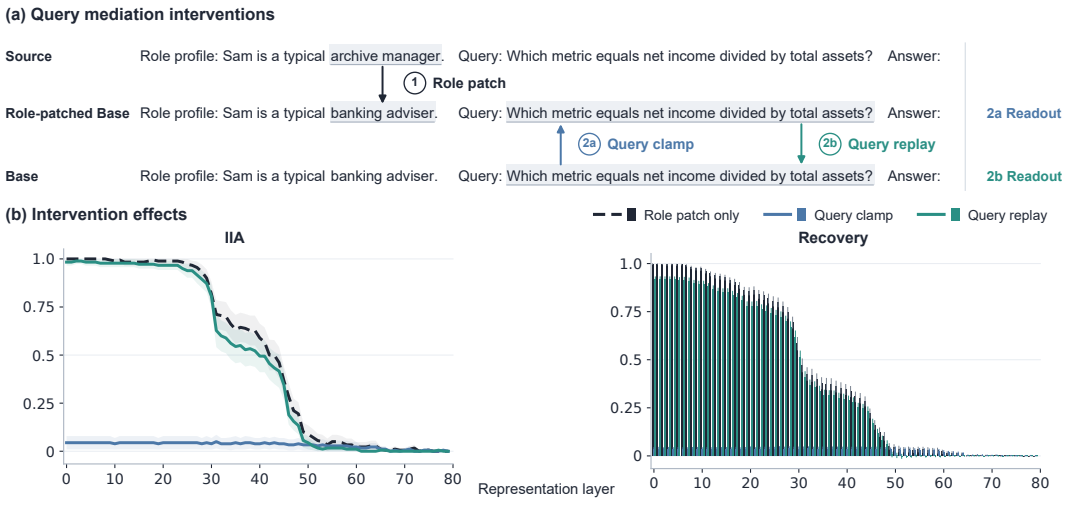

Figure 3: **Query mediation of role-induced behavioral effects.** (a) A Role patch (1) establishes the reference effect. Query clamp (2a) restores the original base Query representations in the Role-patched run, whereas Query replay (2b) transfers the Role-patched Query representations into a base run without Role patching. Clamp and replay are separate conditions, each applied throughout the layers following the Role patch. (b) IIA (left) and normalized logit-margin recovery (right), compared with Role patching alone. The horizontal axis indicates the Role-patch layer.

does not guarantee that a given target model will exhibit the intended behavioral contrast. We therefore qualify families separately for each target model, retaining only those in which both reciprocal pairs exhibit the expected contrast between a direct response and `Unknown`. For each target model, the resulting qualified set contains 90 reciprocal families, corresponding to 180 fixed-query counterfactual pairs across 60 professional roles. Further details on dataset construction and qualification are provided in Appendix C.

3.3 Models and Causal Effect Measures

For the controlled-dataset experiments, we analyze a smaller and a larger checkpoint from each of two model families: Qwen2.5-7B-Instruct and Qwen2.5-72B-Instruct (Yang et al., 2024), and Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct (Llama Team, AI @ Meta, 2024). For readability, the main-text figures consistently show Qwen2.5-72B-Instruct, with corresponding results for the other models reported in Appendix H. Our primary criterion is whether an intervention produces the counterfactual outcome specified by the experiment; interchange intervention accuracy (IIA) reports the proportion of cases in which it does (Geiger et al., 2022). This success criterion does not measure the magnitude of changes in output preference. When such graded effects are informative, we additionally use the experiment-specific output logit margin as a continuous readout (Wang et al., 2023; Zhang & Nanda, 2024). To compare these effects across counterfactual pairs, we express the intervention-induced margin shift as a fraction of the full margin gap between the two matched, unintervened runs. We call this score normalized logit-margin recovery: 0 means that the margin remains at its pre-intervention value, whereas 1 means that it reaches the value observed in the matched counterfactual run. Where shown in results, shaded bands and error bars represent approximate 95% confidence intervals. Further experimental and measurement details, including the construction of these uncertainty estimates, are provided in Appendix F.

4 Mechanistic Analysis of Cognitive Boundaries

Using the controlled counterfactual pairs, we perform layerwise interventions on representations at three concrete semantic anchors in the prompt. As illustrated in Figure 2(a), the Role is anchored in the occupation span of the profile, the Query in the question itself, and the Response at the final input position from which the constrained behavior is read out.

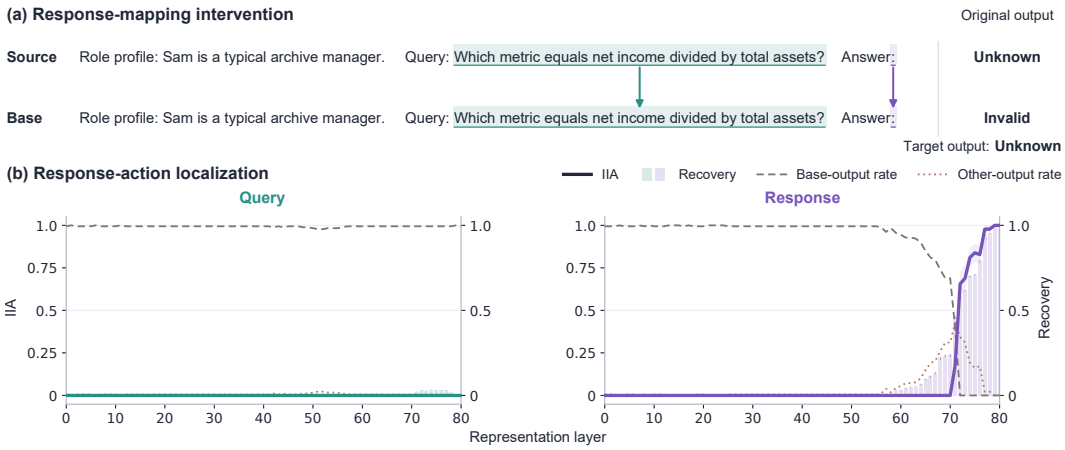

Figure 4: **Localization of response-specific causal control.** (a) The source and base share the same beyond-boundary role–query instance but differ in the instruction-prescribed response: `Unknown` for the source and `Invalid` for the base. The response instructions are omitted from the schematic. Arrows denote alternative interventions at one highlighted position and one representation layer per run. (b) Solid lines show IIA for producing the source-prescribed `Unknown`. Dashed and dotted lines show the rates of retaining `Invalid` and producing other outputs, respectively (left axes). Bars show normalized recovery of the `Unknown`–`Invalid` logit margin (right axes).

4.1 Causal Structure Identification

To identify the causal structure underlying the behavioral contrast, we first locate the representations that contribute to it. The analyses in this subsection are conducted in both counterfactual directions. In the results presented here, the beyond-boundary run serves as the source and the within-boundary run as the base; reverse-direction results are provided in Appendix D. For localization, we patch source representations into the base at one position and one layer at a time within each fixed-query pair, testing whether the response changes to `Unknown` (Figure 2(a)). Figure 2(b) reveals an ordered distribution across model depth: Role-position interventions are effective at earlier layers and weaken with depth, Query-position interventions peak within a narrower intermediate-layer interval, and Response-position interventions become effective at later layers. To test whether this ordering reflects propagation between positions, we intervene at an upstream position and replay the resulting downstream representations in a fresh base run (Appendix D.2). These replay tests recover the effects of Role patching from later Query and Response representations, and those of Query patching from later Response representations, supporting causal transmission across these positions.

Role patching affects both Query and Response representations, but the extent to which Query mediates its behavioral effect remains unresolved. We compare Role patching alone with two complementary interventions to assess this mediation (Figure 3(a)). In *Query clamp*, we retain the Role patch but restore the Query representations at every subsequent layer to their values in the unpatched base run. In *Query replay*, we instead run the base without a Role patch and replay, at those same layers, the Query representations obtained from the Role-patched base. Across layers where Role patching alone is effective, Query clamp reduces IIA to near zero, whereas Query replay largely reproduces the Role-only IIA profile (Figure 3(b)). The recovery results show that clamp leaves little of the original logit-margin effect, while replay recovers most of it. Together with the localization and cross-layer results, these findings support a causal structure in which the effect of Role-position interventions on the response is mediated primarily through subsequent Query representations.

4.2 Causal Computation Identification

The causal structure identified above shows how the effect of changing the role profile propagates toward the response. Within this structure, we further examine how that effect is translated into control over a concrete response. Since the role precedes the query, information about both first becomes jointly available within the Query span; we therefore focus on the Query and Response positions.

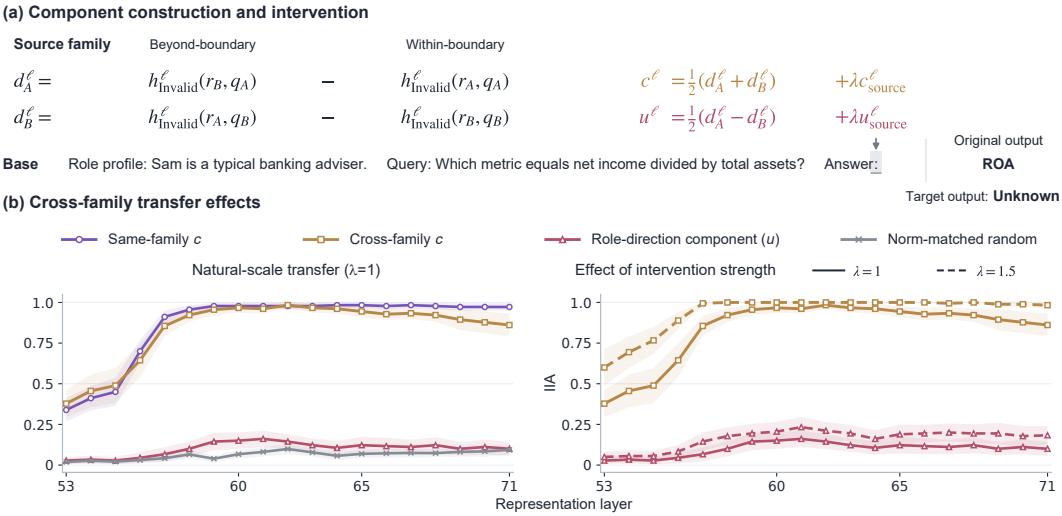

Figure 5: **Cross-family transfer of Response components.** (a) Components are constructed under the `Invalid` response rule and transferred to within-boundary bases under the `Unknown` rule. Each run adds one component at a single Response layer, with subsequent computation proceeding normally. (b) IIA for producing the base-prescribed `Unknown`. Left: same-family and cross-family c^ℓ, cross-family u^ℓ, and random directions matched to the source c^ℓ in norm, at the original scale ($\lambda = 1$). The same-family reference uses the base's own family to construct c^ℓ. Right: cross-family c^ℓ and u^ℓ at $\lambda = 1$ (solid) and $\lambda = 1.5$ (dashed).

To locate where this effect becomes tied to a concrete response, we hold each beyond-boundary role–query instance fixed and change only the instruction-prescribed response from `Unknown` to `Invalid`. The two runs express the same withholding behavior through different prescribed responses. We patch representations from the `Unknown`-mapped source into the `Invalid`-mapped base at one position and one layer at a time, testing whether the output switches from `Invalid` to `Unknown` (Figure 4(a)). The reversed mapping is also tested, with results reported in Appendix D.4. Query interventions rarely transfer the source response, whereas Response interventions do so only at late layers (Figure 4(b)). This contrasts with Figure 2(b), where Query and earlier Response interventions already change boundary-dependent behavior, suggesting a separation between the effect on withholding and the choice of response used to express it. We further test whether these earlier behavioral effects persist when patching from a beyond-boundary source into a within-boundary base with a different response rule (Appendix D.5). These cross-mapping results show that interventions at earlier Response layers still induce withholding according to the base response rule, whereas later interventions transfer the source-prescribed response. This layerwise shift, accompanied by a transient dip in overall withholding efficacy, supports a *transformation window* in which the role-induced behavioral effect is translated into control over the response prescribed by the instruction.

Within this transformation, we further examine what this earlier behavioral effect reflects. Complete Response patches transfer the entire state, so we utilize matched representation differences to retain what changes with the role profile (Rimsky et al., 2024). Let $h^\ell_{\text{Invalid}}(r, q)$ denote the Response representation at layer ℓ under the instruction prescribing `Invalid` for beyond-boundary queries. For each fixed query, we subtract the within-boundary representation from its matched beyond-boundary counterpart:

$$\begin{aligned} d_A^\ell &= h^\ell_{\text{Invalid}}(r_B, q_A) - h^\ell_{\text{Invalid}}(r_A, q_A), \\ d_B^\ell &= h^\ell_{\text{Invalid}}(r_A, q_B) - h^\ell_{\text{Invalid}}(r_B, q_B). \end{aligned} \tag{1}$$

Within each pair, the contribution of the query alone cancels, leaving the role contrast and changes in the joint contribution of the role and query. We use the reversal of the role contrast across the two pairs to form

$$c^\ell = \frac{d_A^\ell + d_B^\ell}{2}, \qquad u^\ell = \frac{d_A^\ell - d_B^\ell}{2}. \tag{2}$$

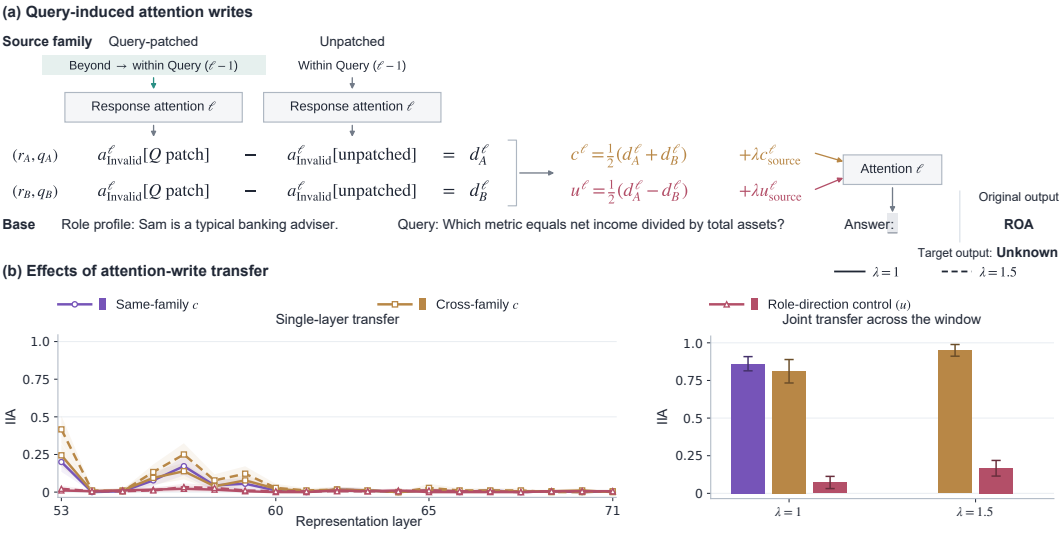

Figure 6: **Query-induced attention writes at the Response position.** (a) $a^{\ell}_{\text{Invalid}}$ denotes the attention output at the Response position under the `Invalid` response rule. Differences between Query-patched and unpatched outputs yield d^{ℓ}_A and d^{ℓ}_B, from which c^{ℓ} and u^{ℓ} are constructed and added in separate base runs. (b) IIA for producing the base-prescribed `Unknown`. Cross-family c^{ℓ} and u^{ℓ} are tested at the original ($\lambda = 1$) and amplified ($\lambda = 1.5$) scales, with same-family c^{ℓ} at the original scale as the reference. Left: independent single-layer additions, with solid and dashed curves denoting the original and amplified scales, respectively. Right: joint additions of the layer-specific components across the selected interval.

The half-sum c^{ℓ} cancels the role contrast, retaining the average change in the joint contribution. The half-difference u^{ℓ} instead counteracts the shared change in the joint contribution, retaining the role contrast as a control (Figure 5(a)). Adding either complete difference to its corresponding within-boundary state would recover the matched beyond-boundary state. Since c^{ℓ} and u^{ℓ} each contain part of these complete differences, same-family additions could partly reproduce the original patching effect. We therefore match each base family to a source family sharing neither roles nor queries with it. Each component is added in a separate run to a within-boundary base at one Response layer within the interval selected from the cross-mapping results, testing whether the response changes to `Unknown`, the withholding response prescribed by the base instruction. To assess the selectivity of cross-family transfer, we compare cross-family c^{ℓ} with a same-family c^{ℓ} reference and two controls: cross-family u^{ℓ} and random directions matched to the source c^{ℓ} in norm. Both cross-family components are tested at their original scale and at 1.5 times that scale. Detailed derivations and the full intervention protocol are provided in Appendix D.6. At the original scale, cross-family c^{ℓ} induces `Unknown` at rates close to the same-family reference over most selected layers, whereas both controls yield low rates (Figure 5(b), left). Increasing the scale strengthens its effect, particularly at earlier layers, while u^{ℓ} remains much less effective (Figure 5(b), right). These results show that the boundary-related component retained in c^{ℓ} remains effective even when both the role and query change. We interpret this transferable component as a shared *conflict signal*. Together with the response-mapping results, this selective transfer provides causal evidence for an internal process in which the model assesses whether a query conflicts with the role's cognitive boundary and uses this signal to constrain its response.

4.3 Causal Computation Reconstruction

Having identified this computation, we next reconstruct how it is formed. Since the role-induced effect is mediated primarily through Query (Section 4.1), we begin by examining what attention at Response writes when reading Query. Because attention at the Response position can also read the representation already present there, we isolate the effect of the Query representation by patching only the within-boundary Query representation at layer $\ell - 1$ with its matched beyond-boundary counterpart and recording the immediately following attention output at the Response position in

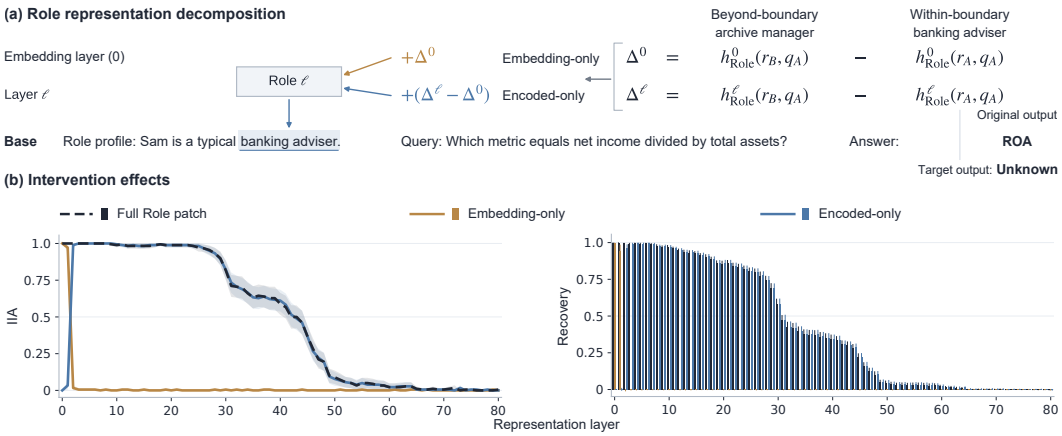

Figure 7: **Causal effects of Role representation components.** (a) Embedding-only and Encoded-only add Δ^0 and $\Delta^\ell - \Delta^0$, respectively, to the base Role representation at layer ℓ, in separate runs. (b) IIA (left) and normalized logit-margin recovery (right), compared with full Role patching. The horizontal axis indicates the intervention layer for both conditions. At layer 0, Embedding-only coincides with full Role patching, and the Encoded-only difference is zero.

layer ℓ (Figure 6(a)). To test whether these writes contain the conflict signal identified above, we apply the decomposition in Equation 2 to the differences between patched and unpatched attention outputs under the `Invalid` rule. The resulting c^ℓ and u^ℓ are added separately to attention outputs in within-boundary base runs under the Unknown rule, retaining the family pairings and scales used in Section 4.2. We compare single-layer additions with joint additions of the layer-specific components across the same selected interval; full derivations and intervention details are provided in Appendix D.7. At the original scale, single-layer additions of cross-family c^ℓ already induce Unknown at several layers near the start of the selected interval, whereas joint addition produces a much stronger effect close to the same-family reference (Figure 6(b)). Increasing the scale strengthens the effect of c^ℓ, while u^ℓ remains much weaker. These results support a mechanism in which, when the query lies beyond the role's cognitive boundary, attention writes a conflict signal from Query into the Response representation. Multiple writes jointly reinforce this signal, which is subsequently mapped to the response prescribed by the instruction.

We next follow the same causal structure upstream to examine the role information supplied to Query. We compare the downstream effects of the profile's initial representation with those of the additional role information encoded at the Role position over subsequent layers. Let $h^\ell_{\text{Role}}(r, q)$ denote the Role representation at layer ℓ in the run with role profile r and query q. For each fixed-query pair, we denote the difference between the source and unpatched base Role representations by Δ^0 at the embedding layer and by Δ^ℓ at layer ℓ. In separate runs at the same layer ℓ, we add either the initial difference Δ^0 (*Embedding-only*) or the difference formed after the embedding layer, $\Delta^\ell - \Delta^0$ (*Encoded-only*), to the base Role representation. Full Role patching provides the reference for both conditions (Figure 7(a)). We evaluate both counterfactual directions; the figure shows a beyond-boundary source and a within-boundary base, with reverse-direction results, detailed derivations, and intervention procedures provided in Appendix D.8. Beyond the initial layers, Encoded-only closely reproduces the IIA and normalized logit-margin recovery of full Role patching over its effective range, whereas Embedding-only has little effect (Figure 7(b)). Together with the Query-mediation results, these findings support a mechanism in which the role profile is encoded at the Role position into a role state that causally influences subsequent processing.

4.4 Mechanistic Synthesis

Taken together, our analyses at three successive levels support the four-stage process in Figure 1. In *causal structure identification*, we trace how the role profile's effect propagates from Role to Response, primarily through Query (Section 4.1). In *causal computation identification*, we identify a shared conflict signal that is translated into control over a concrete response within a later transformation window (Section 4.2). In *causal computation reconstruction*, we work backward along this

structure, showing how attention reads Query to write and reinforce the conflict signal at Response and how the encoded role state conditions Query processing (Section 4.3). Repeating all intervention analyses on a role-disjoint confirmation set yields qualitatively consistent results (Appendix I).

5 External Validation on TimeChara

To assess whether this mechanistic account extends beyond our controlled setting, we evaluate it on TimeChara (Ahn et al., 2024). We use its *future* subset, where questions concern events in the character's narrative, but the specified time point places those events beyond the character's current cognitive boundary. We retain the original prompt format and standardize withholding as a prescribed one-word response, leaving direct answers open-ended. To disrupt the model's use of this temporal constraint while preserving token alignment, we follow Meng et al. (2022) and add Gaussian noise only to the embeddings of the character's time-setting tokens. We retain cases in which the perturbation changes withholding to direct responding and use the paired runs to test whether internal interventions restore withholding. We repeat this construction with three random seeds for Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct. The key causal signatures identified in the controlled setting also recur in both models (Table 1). Full details are provided in Appendix E.

Table 1: **External validation results on TimeChara.** The four analyses correspond to Figures 2, 3, 4, and 5, respectively. Layer entries list the layers where IIA averaged across the three seeds exceeds 0.5; — denotes none. Relative IIA is the per-seed ratio of summed condition to reference IIA over all evaluated layers for Query mediation and the selected Response layers for conflict transfer, reported as mean $\pm$ SD across seeds.

Analysis	Comparison	Llama-3.1-70B	Qwen2.5-72B
Causal effect localization *Layer entries*	Role	0–17	0–48
	Query	30	46–56
	Response	34–80	58–80
Query mediation *Relative IIA*	Replay / Role reference	0.930 ± 0.009	1.015 ± 0.009
	Clamp / Role reference	0.118 ± 0.032	0.065 ± 0.007
Response mapping *Layer entries*	Query	—	—
	Response	76–80	77–80
Conflict transfer *Relative IIA*	Cross-context / same-context	0.841 ± 0.010	1.018 ± 0.046
	Residual control / same-context	0.155 ± 0.015	0.189 ± 0.036

6 Discussion and Limitations

Our findings show that the role profile shapes response formation by influencing how the current query is processed. Although our analyses establish Query as the primary mediator of this influence, they do not yet resolve how the role state and query content are combined to form the role-conditioned query state. The location of this interaction is also tied to the input order, since our profile-before-query setting makes role information available during Query processing. Changing this order may therefore alter where the two sources of information are combined. Whether the role's cognitive boundary still constrains responses through a conflict signal under such changes remains to be tested.

7 Conclusion

In this work, we provide a representation-level causal account of how language models produce responses that respect role-specific cognitive boundaries. The model uses the role profile to condition the processing of the current query. When the query lies beyond the role's cognitive boundary, this processing gives rise to a shared conflict signal that is subsequently transformed into concrete response behavior. This account helps explain how the same model can possess broad parametric knowledge yet answer from the narrower cognitive perspective specified by a role.

AI USE STATEMENT

We used LLMs to assist with the writing and polishing of the manuscript, including wording, readability, and the flow of the text. For the controlled dataset, we used Qwen2.5-72B-Instruct to generate candidate queries and reference answers from ESCO occupation–knowledge associations. The generated candidates underwent format checks, semantic validation with Llama-3.1-70B-Instruct, and model-specific qualification before inclusion in the experiments. These procedures are detailed in Appendix C, with prompt templates provided in Appendix G.1. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

This study uses pretrained language models, ESCO, and TimeChara in accordance with their applicable licenses and terms of use. No human-participant experiments were conducted, and no personal data were collected for this study. The controlled role profiles define experimental conditions and should not be interpreted as assessments of the knowledge or professional competence of real individuals.

REPRODUCIBILITY STATEMENT

We provide the generated datasets and code for data construction and intervention experiments at `https://anonymous.4open.science/r/role-cognitive-boundaries-3BC8`. The paper and appendices describe how these datasets are constructed and used in the experiments, with implementation and measurement details in Appendix F and full prompt templates in Appendix G.

REFERENCES

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3291–3325, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.197. URL https://aclanthology.org/2024.findings-acl.197/.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14638–14663. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.785. URL https://aclanthology.org/2024.acl-long.785/.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

European Commission. ESCO v1.2.1: European skills, competences, qualifications and occupations. European Skills, Competences, Qualifications and Occupations (ESCO), December 2025. URL https://esco.ec.europa.eu/en/about-esco/escopedia/escopedia/esco-versions. Version 1.2.1.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9d1b7fc578c0d2d6431fc26d736ecaf3-Abstract-Conference.html.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1828–1843. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.144. URL https://aclanthology.org/2021.acl-long.144/.

Zhenpeng Gao, Xiaofen Xing, and Xiangmin Xu. TailorRPA: A retrieval-based framework for eliciting personalized and coherent role-playing agents in general domain. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5381–5412, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.288. URL https://aclanthology.org/2025.findings-emnlp.288/.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338. PMLR, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Sunkyung Han, Nahyeon Park, Gaeun Seo, Seunghyun Yoon, and JinYeong Bak. CHARM: Character hallucination for multicultural role play benchmark. arXiv preprint arXiv:2609.01352, 2026. URL https://arxiv.org/abs/2609.01352.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.470. URL https://aclanthology.org/2024.acl-long.470/.

International Labour Office. *International Standard Classification of Occupations: ISCO-08. Volume 1: Structure, Group Definitions and Correspondence Tables*. International Labour Office, Geneva, 2012. URL https://researchrepository.ilo.org/esploro/outputs/book/International-Standard-Classification-of-Occupations-ISCO-08/995673746502676.

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*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Sheng Liu, Haotian Ye, Lei Xing, and James Y. Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32287–32307. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24bx.html.

Wenhao Liu, Siyu An, Junru Lu, Muling Wu, Tianlong Li, Xiaohua Wang, Changze Lv, Xiaoqing Zheng, Di Yin, Xing Sun, and Xuanjing Huang. Tell me what you don't know: Enhancing refusal capabilities of role-playing agents via representation space analysis and editing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5983–6005, Vienna, Austria, July

2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.311. URL https://aclanthology.org/2025.findings-acl.311/.

Llama Team, AI @ Meta. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. RoleMRC: A fine-grained composite benchmark for role-playing and instruction-following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21008–21030, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1082. URL https://aclanthology.org/2025.findings-acl.1082/.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7828–7840, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.423. URL https://aclanthology.org/2024.acl-long.423/.

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, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4ffbfbcf4ad7304d57158b046525e46c-Abstract-Conference.html.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis. *Computational Linguistics*, 52(1):331–378, 2026. doi: 10.1162/coli.a.572. URL https://aclanthology.org/2026.cl-1.10/.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6gO6KTRMpG.

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, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, and Julian McAuley. Mitigating hallucination in fictional character role-play. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14467–14479, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.846. URL https://aclanthology.org/2024.findings-emnlp.846/.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pp. 13153–13187. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.emnlp-main.814/.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4ae163cb8788970e53b4fd9578141139-Abstract-Conference.html.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.638. URL https://aclanthology.org/2024.acl-long.638/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Kai Wang, Haoyang You, Yang Zhang, and Zhongjie Wang. Memory-driven role-playing: Evaluation and enhancement of persona knowledge utilization in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23475–23510, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1175. URL https://aclanthology.org/2026.findings-acl.1175/.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Noah Wang, Z.Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878. URL https://aclanthology.org/2024.findings-acl.878/.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1840–1873, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.102. URL https://aclanthology.org/2024.acl-long.102/.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, Qiang Qu, Feiteng Fang, Min Yang, Lianli Gao, Jingkuan Song, Fei Huang, and Yongbin Li. OmniCharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26318–26331, Vienna, Austria, July 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1276. URL https://aclanthology.org/2025.acl-long.1276/.

Wenyuan Zhang, Shuaiyi Nie, Jiawei Sheng, Zefeng Zhang, Xinghua Zhang, Yongquan He, and Tingwen Liu. Revealing and mitigating the challenge of detecting character knowledge errors in LLM role-playing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33279–33302, Suzhou, China, November 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1689. URL https://aclanthology.org/2025.emnlp-main.1689/.

A Related Work

In role-playing, LMs are expected to reflect the characteristics of the role they portray, including its speaking style, personality, and knowledge (Wang et al., 2024a;b; Tu et al., 2024). A particular challenge is that the model may know the answer to a query that lies beyond the role's cognitive boundary (Shao et al., 2023). TimeChara examines this issue by placing characters at specified points in a story and checking whether their responses reveal knowledge of later events (Ahn et al., 2024). RoleMRC also considers the abilities specified in a role profile, testing whether models answer questions the role can answer and follow instructions to decline or attempt questions beyond those abilities (Lu et al., 2025). CHARM separately tests whether models explicitly identify questions beyond a character's knowledge and whether they withhold answers to those questions when responding in character (Han et al., 2026).

Alongside these evaluations, existing work has developed methods to improve models' adherence to role-specific cognitive boundaries. RoleFact checks factual claims in a draft role-playing response against retrieved story evidence and rewrites the response to remove claims that remain insufficiently supported (Sadeq et al., 2024). TailorRPA incorporates both questions related to a character's experiences and questions outside the character's knowledge into instruction-tuning data, teaching the model when to answer and when to express ignorance (Gao et al., 2025). In addition, Liu et al. (2025) observe rejection and direct-response regions within each character's cluster using t-SNE on last-layer representations, and then apply representation editing to encourage refusal of conflicting queries.

These approaches improve models' adherence to role-specific cognitive boundaries, but leave open how this behavior is produced internally. In a related belief-tracking task, Prakash et al. (2026) use causal interventions to reconstruct how LMs retrieve what a particular character believes about the state of an object. Our work addresses a related question in role-playing: how a model uses the current role profile to determine whether the knowledge requested by a query lies beyond the role's cognitive boundary and to produce a response consistent with that boundary.

B Intervention-Based Causal Analysis

Intervention-based causal analysis examines how controlled changes to model variables affect subsequent computation and behavior (Mueller et al., 2026). This section focuses on interventions on internal representations, covering activation patching, causal mediation analysis, and activation steering.

B.1 Activation Patching

Activation patching tests the causal contribution of an internal activation by setting its value in a base run with the corresponding value from a source run (Meng et al., 2022; Zhang & Nanda, 2024; Arora et al., 2024). Let x_{b} and x_{s} denote matched base and source inputs, and let $\mathcal{M}(x)$ denote the output of a fixed model on input x. We write $h_p^{\ell}(x)$ for the selected activation at layer ℓ and position

p. For a replacement value a, the notation $\mathcal{M}(x; h_p^\ell \leftarrow a)$ denotes the output obtained when this activation is set to a. The unpatched and patched outputs are

$$o_{\mathrm{b}} = \mathcal{M}(x_{\mathrm{b}}), \qquad o_{\mathrm{s}} = \mathcal{M}(x_{\mathrm{s}}),$$
$$o_{\mathrm{b}\leftarrow\mathrm{s}}^{\ell,p} = \mathcal{M}\big(x_{\mathrm{b}}; h_p^\ell \leftarrow h_p^\ell(x_{\mathrm{s}})\big)\,. \tag{3}$$

The base input and model parameters remain unchanged, and subsequent computation proceeds from the modified activation.

The intervention is evaluated by comparing $o_{\mathrm{b}\leftarrow\mathrm{s}}^{\ell,p}$ with the unpatched base output o_{b}, using the source output o_{s} as a reference for the behavioral contrast. A shift toward the source behavior under the chosen behavioral measure provides evidence that the selected activation contributes causally to that contrast (Huang et al., 2024; Feng & Steinhardt, 2024). Repeating the intervention across layers and positions identifies where such patching affects model behavior.

B.2 Causal Mediation Analysis

Activation patching at separate locations can establish local causal relevance without determining the causal relations among those locations (Conmy et al., 2023; Merullo et al., 2024). Causal mediation analysis examines the extent to which the behavioral effect of an upstream intervention is transmitted through an intermediate representation, called a mediator (Vig et al., 2020; Finlayson et al., 2021; Todd et al., 2024).

For illustration, consider patching the activation at layer ℓ_u and position p_u in the base run with its source value. We test whether the resulting behavioral effect passes through the activation at a later layer $\ell_m > \ell_u$ and position p_m. First, we ask what effect the upstream patch has when this mediator remains at its unpatched base value. We therefore keep the upstream patch and patch the mediator back to $h_{p_m}^{\ell_m}(x_{\mathrm{b}})$, producing

$$o_{\mathrm{direct}} = \mathcal{M}\big(x_{\mathrm{b}}; h_{p_u}^{\ell_u} \leftarrow h_{p_u}^{\ell_u}(x_{\mathrm{s}}),\ h_{p_m}^{\ell_m} \leftarrow h_{p_m}^{\ell_m}(x_{\mathrm{b}})\big)\,. \tag{4}$$

Comparing o_{direct} with o_{b} under the chosen behavioral measure evaluates the effect of the upstream patch while the mediator is held fixed. This is a direct effect relative to the selected mediator.

We next ask whether the mediator value produced by the upstream patch can affect behavior without applying that patch in the test run. Let $\widetilde{h}_{p_m}^{\ell_m}$ denote the mediator activation recorded from a base run with only the upstream patch. In a separate base run, we leave the upstream activation unpatched and patch only the mediator with this recorded value, giving

$$o_{\mathrm{indirect}} = \mathcal{M}\Big(x_{\mathrm{b}}; h_{p_m}^{\ell_m} \leftarrow \widetilde{h}_{p_m}^{\ell_m}\Big)\,. \tag{5}$$

Comparing o_{indirect} with o_{b} under the same behavioral measure evaluates the effect carried by the induced mediator value. This is an indirect effect of the upstream patch through the selected mediator (Vig et al., 2020). Together, these comparisons assess the extent to which the selected representation mediates the behavioral effect of the upstream patch.

B.3 Activation Steering

Patching a complete source activation tests its overall effect at the selected location. To examine the behavioral effect of the change between two matched runs, we can compute the difference between their activations and add it to an activation in a base run. In this form of activation steering, the added difference serves as the steering vector (Rimsky et al., 2024; Arditi et al., 2024).

Let d_p^ℓ denote the activation difference used for the intervention. For a single pair of matched inputs x^+ and x^- associated with the target and contrasting behaviors, respectively, it is

$$d_p^\ell = h_p^\ell(x^+) - h_p^\ell(x^-). \tag{6}$$

When N matched pairs $\{(x_i^+, x_i^-)\}_{i=1}^N$ are used, the intervention difference can instead be formed by averaging their corresponding activation differences to emphasize changes that are consistent across pairs:

$$d_p^\ell = \frac{1}{N}\sum_{i=1}^{N}\big[h_p^\ell(x_i^+) - h_p^\ell(x_i^-)\big]\,. \tag{7}$$

For a base input x_{b}, the difference is added to the selected activation:

$$o_{\mathrm{b}}^{\ell,p}(\lambda) = \mathcal{M}\big(x_{\mathrm{b}}; h_p^{\ell} \leftarrow h_p^{\ell}(x_{\mathrm{b}}) + \lambda d_p^{\ell}\big)\,. \tag{8}$$

The coefficient λ controls the sign and scale of the addition, with $\lambda = 0$ recovering the unpatched base output o_{b}. The input and model parameters remain fixed, and subsequent computation proceeds from the modified activation (Liu et al., 2024). Comparing $o_{\mathrm{b}}^{\ell,p}(\lambda)$ with o_{b} under the chosen behavioral measure assesses the causal effect of the added difference on the target behavior (Li et al., 2023).

C Controlled Dataset Construction and Qualification

This appendix details the construction of the controlled dataset introduced in Section 3.1. We first generate a role–query candidate pool from occupation–knowledge associations, then use semantic validation and model-specific qualification to select queries and form counterfactual families for analysis.

C.1 Role Selection and Query Generation

ESCO (European Skills, Competences, Qualifications and Occupations) is a classification developed by the European Commission that describes occupations, skills, and knowledge concepts, including explicit links between occupations and their knowledge requirements (European Commission, 2025). Its occupations are mapped to the International Standard Classification of Occupations (ISCO-08), whose major and sub-major groups are used here to organize role selection. We use the English v1.2.1 release of ESCO and focus on Major Group 2, *Professionals*, whose occupations correspond to ISCO skill level 4 and require theoretical and factual knowledge in specialized fields (International Labour Office, 2012). Within this group, individual ESCO occupations serve as roles, and their sub-major groups are used for subsequent pairing across professional fields. Differences in the professional knowledge expected of these roles provide a basis for constructing candidate pairs in which the same query lies within one role's cognitive boundary and beyond another's. For each occupation, ESCO identifies *essential* knowledge that is normally required across employers and work contexts. These existing associations guide the generation of questions that a typical practitioner would be expected to answer.

To focus on knowledge tied to particular professional backgrounds, we retain essential knowledge items marked by ESCO as *occupation-specific* or *sector-specific*. For each retained item, we ask Qwen2.5-72B-Instruct (Yang et al., 2024) to generate candidate queries and reference answers from the occupation name and the item's label and description. The model is instructed to generate factual questions that test the associated knowledge and admit a uniquely determined one-word answer, without naming the occupation or revealing the answer in the query. Candidates that pass basic format checks are stored with their reference answers and associated occupations and knowledge items, forming a role–query candidate pool for subsequent validation and pairing. To leave alternatives for these later stages, the pool contains up to 100 candidate occupations per sub-major group, with a final target of 10 roles per group. Each candidate role must have at least 3 eligible knowledge items, and 10 candidate query–answer pairs are requested for every retained item.

C.2 Semantic Validation and Model-Specific Qualification

We use Llama-3.1-70B-Instruct (Llama Team, AI @ Meta, 2024) to check that each candidate query directly tests the associated essential knowledge, that a typical practitioner in the corresponding occupation would normally be expected to know the answer, and that its reference answer is correct and uniquely determined. After retaining candidates that pass these checks, we keep roles with validated queries from at least three distinct knowledge items for subsequent pairing. Table 2 reports the shared candidate pool and the queries and roles retained after semantic validation.

To construct counterfactual pairs in which the same query lies within one role's cognitive boundary and beyond another's, we pair roles from different sub-major groups and combine each role's validated queries with both profiles. These pairs are prepared under the original instruction prescribing `Unknown` and an alternative prescribing `Invalid` for withholding, as required by the response-

Table 2: **Shared candidate pool and semantic validation.** Query candidates are format-valid query–answer pairs. Percentages give the proportion retained after semantic validation. Eligible roles have validated queries from at least three distinct knowledge items.

ISCO-08 sub-major group	Candidate roles	Query candidates	Validated candidates	Eligible roles
21 Science and engineering professionals	100	5,860	4,517 (77.1%)	100
22 Health professionals	47	2,672	2,079 (77.8%)	46
23 Teaching professionals	35	1,572	1,261 (80.2%)	35
24 Business and administration professionals	37	1,326	1,031 (77.8%)	37
25 Information and communications technology professionals	67	4,107	3,156 (76.8%)	67
26 Legal, social and cultural professionals	34	1,382	1,024 (74.1%)	33
Total	**320**	**16,919**	**13,068 (77.2%)**	**318**

mapping analysis in Section 4.2.[2] For each target model, we first verify that the paired inputs are token-aligned at the Role, Query, and Response positions under both instructions. We then check the intended behavioral contrast for queries from each role: under both instructions, a qualifying query must elicit a direct response under its associated role and the prescribed withholding response under the other.

For each candidate role pair, we search for one qualifying query from each of three distinct knowledge items per role. The role pair is accepted only when both roles meet this requirement; the two sets of queries are then matched one-to-one to form three reciprocal families. We continue selecting role pairs without reusing roles until each of the six sub-major groups contributes 10 roles. The resulting set for each target model contains 60 roles, 90 reciprocal families, and 180 fixed-query counterfactual pairs. Table 3 reports the token-aligned role pairs available for this selection and the number examined before the target size is reached.

Table 3: **Model-specific role-pair qualification.** Token-aligned role pairs are drawn from 40,418 unordered pairs of eligible roles from different sub-major groups. The last column counts pairs evaluated for behavioral qualification until all group quotas are met. Evaluation counts are reconstructed from the deterministic search order, assuming that each final family set is complete.

Target model	Token-aligned role pairs	Role pairs evaluated
Qwen2.5-7B-Instruct	11,275	41
Qwen2.5-72B-Instruct	11,275	34
Llama-3.1-8B-Instruct	11,275	108
Llama-3.1-70B-Instruct	11,275	32

D Extended Mechanistic Analyses on the Controlled Dataset

This appendix provides additional intervention details and results for the controlled-dataset analyses in Section 4. As in the main text, figures show Qwen2.5-72B-Instruct; results for the other models are reported in Appendix H.

D.1 Causal Effect Localization

We repeat the localization comparison in Figure 2 with the within-boundary run as the source and the beyond-boundary run as the base. The interventions now test whether patching the Role, Query, or Response representation restores direct responding to the same query. Figure 8 shows the same overall ordering as the main-text comparison: Role interventions are effective at earlier layers, Query effects peak in an intermediate interval, and Response effects become pronounced at later layers.

[2] We choose `Invalid` to preserve token alignment between the two instruction versions for all target models.

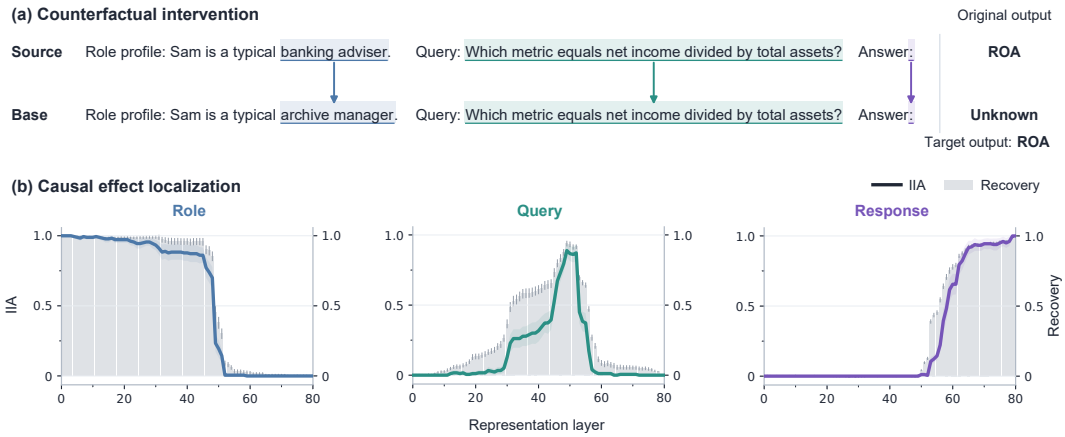

Figure 8: **Reverse-direction causal effect localization.** (a) The within-boundary run is the source and the beyond-boundary run is the base, with direct responding as the target behavior. The arrows indicate separate interventions at the highlighted positions, each performed at one representation layer per run. (b) Solid curves show IIA (left axes), and bars show normalized logit-margin recovery (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

D.2 Cross-Site Representation Replay

To test whether the ordering in Figure 2(b) reflects propagation between positions, we examine whether the effect of an upstream patch can be recovered by replaying a downstream representation. For each fixed-query pair, we patch the source representation at position U and layer i into the base, then let the run continue to a later layer $j > i$ and record the representation at a downstream position V. We replay this recorded representation in a fresh base run, without the upstream patch. Writing $h_U^\ell(x)$ for the activations at position U and layer ℓ, the recorded representation and replay output are

$$\begin{aligned} \widetilde{h}_V^j &= h_V^j\left(x_{\mathrm{b}}; h_U^i \leftarrow h_U^i(x_{\mathrm{s}})\right), \\ o_{U\to V}^{i,j} &= \mathcal{M}\left(x_{\mathrm{b}}; h_V^j \leftarrow \widetilde{h}_V^j\right). \end{aligned} \tag{9}$$

We apply this procedure to Role-to-Query, Query-to-Response, and Role-to-Response replay in both counterfactual directions.

We focus the replay tests on the effective layer regions identified by localization, with results shown in Figures 9 and 10. In both directions, the effect of Role patching can be recovered from later Query and Response representations, while the effect of Query patching can be recovered from later Response representations. These results support propagation of the localized causal effects between Role, Query, and Response.

D.3 Query Mediation

We repeat the Query-mediation comparison in Figure 3 with a within-boundary source and a beyond-boundary base, testing whether Query also mediates the restoration of direct responding. Figure 11 shows that, over the layers where Role patching is effective, Query replay closely reproduces the Role-only IIA and recovery profiles, whereas Query clamp reduces IIA to near zero and leaves little of the logit-margin effect. This supports Query as the primary mediator of the Role intervention's behavioral effect in the reverse direction as well.

D.4 Response Mapping

We reverse the response-mapping comparison in Figure 4, keeping the same beyond-boundary role–query instance but using an `Invalid`-mapped source and an `Unknown`-mapped base. Patching now tests whether the output changes from `Unknown` to `Invalid`. As in the main text, Query interventions

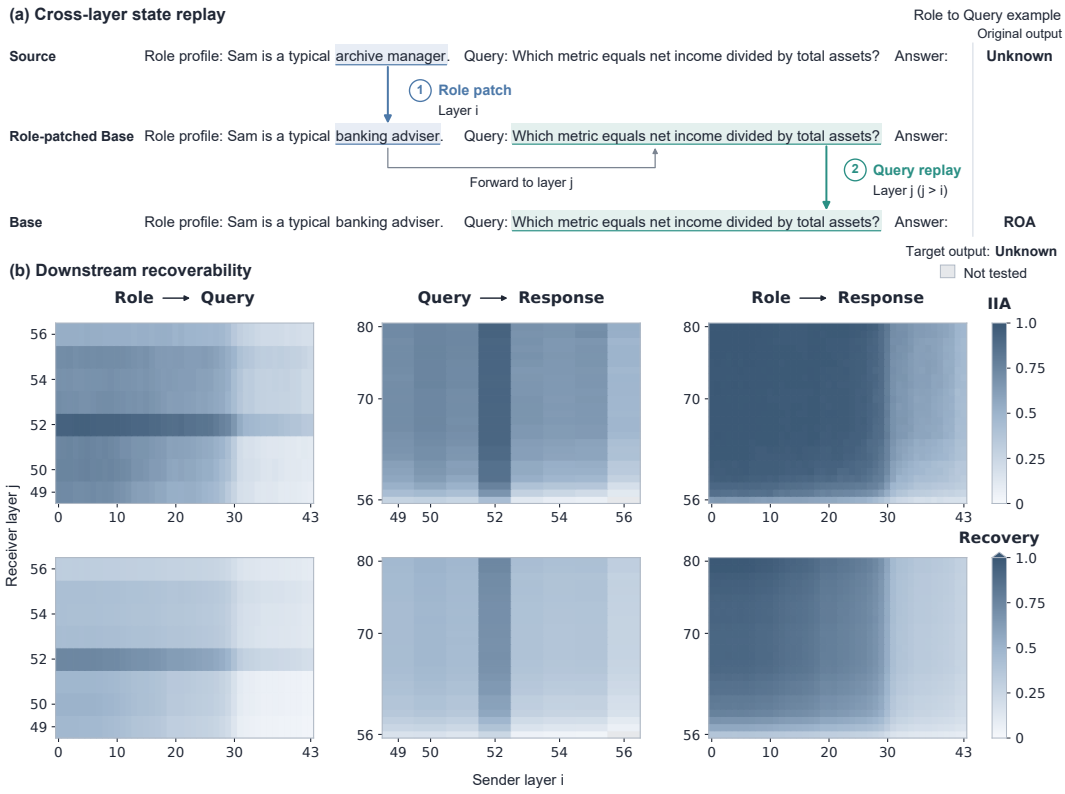

Figure 9: **Cross-site replay with a beyond-boundary source.** (a) Role-to-Query example: after the Role patch at layer i, the Query representation is recorded at layer $j > i$ and replayed in a fresh within-boundary base run (bottom row), without retaining the Role patch. The target behavior is withholding with `Unknown`. (b) Columns identify the sender and receiver positions; horizontal and vertical axes give the sender layer i and receiver layer j, respectively. Top heatmaps show IIA, and bottom heatmaps show normalized logit-margin recovery. Grey cells are not tested. The upper extension of the Recovery color bar denotes values above 1.

rarely transfer the source-prescribed response, whereas Response interventions do so at late layers (Figure 12).

D.5 Cross-Mapping Interventions

In the task defined in Section 2, withholding is the role-consistent behavior for a query beyond the role's cognitive boundary. The distinction examined here is between the boundary-related effect on withholding and control over the particular response used to express it. To test whether this effect persists across response mappings, we patch Query or Response representations from a beyond-boundary source into a within-boundary base with the same query but a different prescribed withholding response. In Figure 13(a), the source instruction prescribes `Invalid` and the base instruction prescribes `Unknown`. After patching, `Unknown` indicates withholding under the base response rule, while `Invalid` indicates transfer of the source-prescribed response. We repeat this comparison with the two prescribed words exchanged and test both assignments in the reverse direction, using a within-boundary source and a beyond-boundary base.

Consistent with Section 4.2, Query and earlier Response interventions induce withholding mainly under the base response rule, whereas later Response interventions produce the source-prescribed response (Figure 13). The reverse interventions also restore direct responding across the different response rules (Figure 14). In Figure 13(b), total withholding IIA at earlier Response layers nearly coincides with the base-prescribed response rate. We further observe the transient dip in cross-mapping efficacy described in the main text: as the patching layer moves later, total IIA declines with

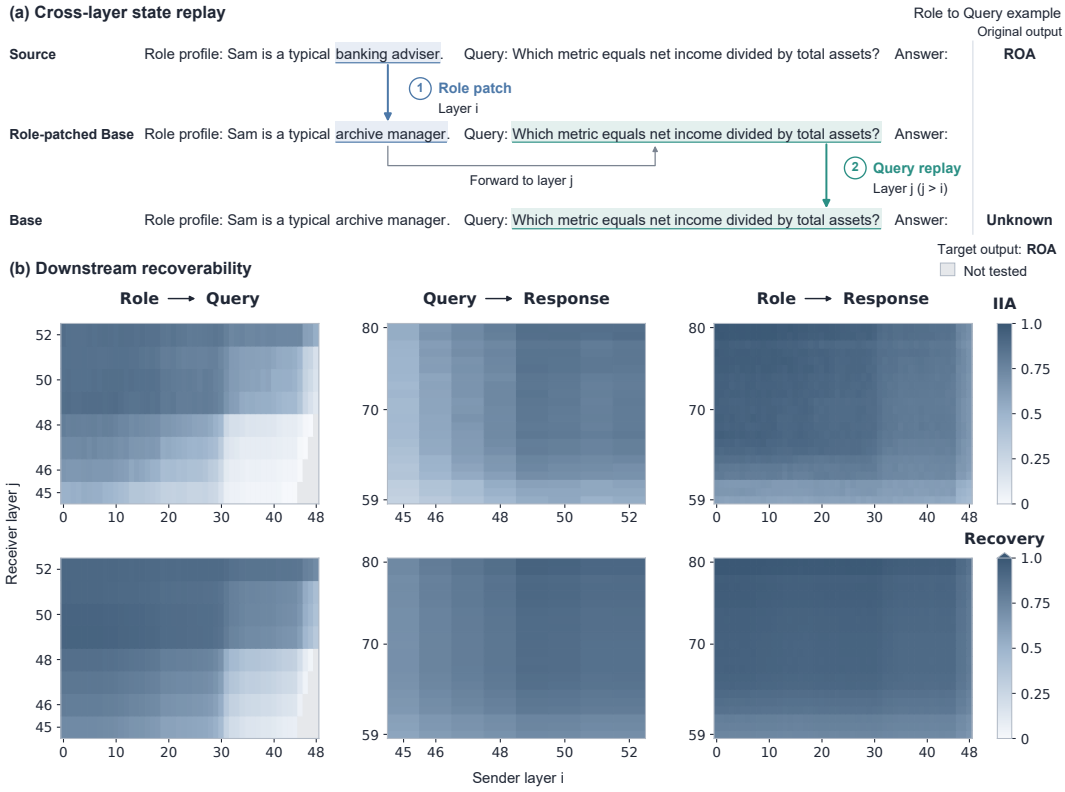

Figure 10: **Cross-site replay with a within-boundary source.** (a) The source and base are reversed relative to Figure 9, so the fresh base is beyond-boundary and the target behavior is direct responding. (b) The heatmaps use the same layout and measures as Figure 9, with layer regions drawn from the reverse-direction localization results. Grey cells are not tested; the upper extension of the Recovery color bar denotes values above 1.

the base-prescribed response rate while the source-prescribed response rate rises. The minimum lies near the intersection of the two response-rate curves. At later layers, total IIA rises again and comes to coincide with the source-prescribed response rate as the base-prescribed response rate approaches zero. This decline and recovery occur with either assignment of the prescribed words but are not observed in the original Response localization (Figure 2) or the reverse interventions (Figure 14). The temporary reduction in efficacy suggests that cross-mapping patches disrupt the conversion to a concrete response while the earlier behavioral effect is weakening and control over the source-prescribed response is not yet fully established. Together with the shift from the base-prescribed response to the source-prescribed response, this pattern strongly supports the transformation window described in the main text.

D.6 Conflict Transfer

This section provides the detailed derivation and intervention protocol for the analysis in Section 4.2. Query mediation shows that the causal effect of Role patching reaches Response primarily through Query. It does not, however, distinguish whether the earlier behavioral effect at Response reflects the separate effects of the role and query or a joint effect arising from their combination. To interpret the matched differences, we assume that these effects contribute to the Response representation through the following decomposition:

$$h^{\ell}_{\text{Invalid}}(r, q) = R^{\ell}(r) + Q^{\ell}(q) + J^{\ell}(r, q), \tag{10}$$

where $R^{\ell}(r)$ and $Q^{\ell}(q)$ describe the contributions of the role and query alone, and $J^{\ell}(r, q)$ describes their joint contribution. We compute the matched differences from the model's Response representations, subtracting the within-boundary representation from its beyond-boundary counterpart for



Figure 11: **Query mediation in the reverse counterfactual direction.** (a) A within-boundary source provides the Role patch to a beyond-boundary base. Query clamp and Query replay are separate conditions, with Role patching alone as their reference. (b) IIA (left) and normalized logit-margin recovery (right) measure their effects in the direction of direct responding. The representation layer on each horizontal axis is the Role-patch layer; the Query interventions are applied at every subsequent layer. Shaded bands and error bars indicate approximate 95% confidence intervals.



Figure 12: **Response mapping with the prescribed words reversed.** (a) The source and base use the same beyond-boundary role–query instance, with `Invalid` prescribed for the source and `Unknown` for the base. The response instructions are omitted from the schematic; each arrow indicates a separate intervention at one highlighted position and one layer. (b) Solid curves show IIA for producing `Invalid`; dashed and dotted curves show the rates of retaining `Unknown` and producing other outputs, respectively (left axes). Bars show normalized recovery of the `Invalid`–`Unknown` logit margin (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

each fixed query:

$$d_A^\ell = h_{\text{Invalid}}^\ell(r_B, q_A) - h_{\text{Invalid}}^\ell(r_A, q_A),$$
$$d_B^\ell = h_{\text{Invalid}}^\ell(r_A, q_B) - h_{\text{Invalid}}^\ell(r_B, q_B). \tag{11}$$

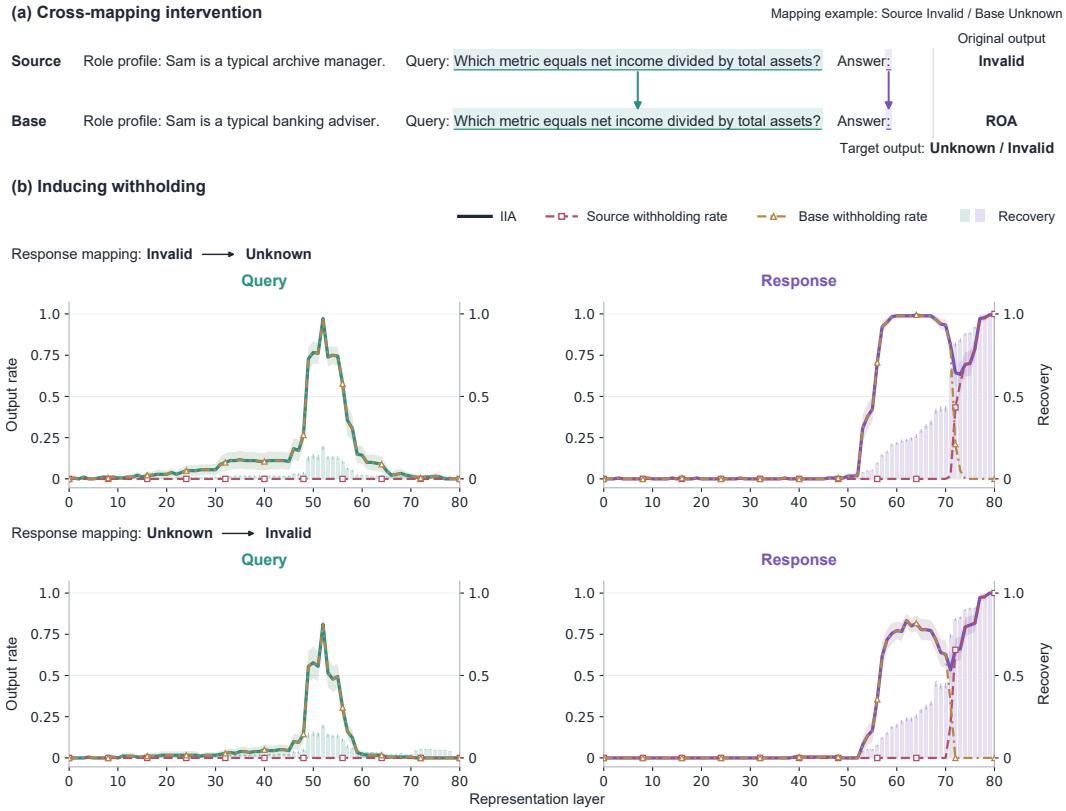

Figure 13: **Cross-mapping interventions that induce withholding.** (a) Query or Response representations from a beyond-boundary source are patched into a within-boundary base, one position and one layer per run. (b) Row labels give the withholding responses prescribed for the source and base, in that order. Solid curves show IIA for producing either prescribed withholding response. Dashed curves show how often the patched base produces the source- and base-prescribed responses; these rates sum to withholding IIA (left axes). Bars show normalized recovery of the logit margin between the source's original withholding output and the base's original direct output (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

Under the assumed decomposition, the two comparisons expand as

$$\begin{aligned} h^{\ell}_{\text{Invalid}}(r_B, q_A) - h^{\ell}_{\text{Invalid}}(r_A, q_A) &= R^{\ell}(r_B) - R^{\ell}(r_A) \\ &\quad + Q^{\ell}(q_A) - Q^{\ell}(q_A) \\ &\quad + J^{\ell}(r_B, q_A) - J^{\ell}(r_A, q_A), \\ h^{\ell}_{\text{Invalid}}(r_A, q_B) - h^{\ell}_{\text{Invalid}}(r_B, q_B) &= R^{\ell}(r_A) - R^{\ell}(r_B) \\ &\quad + Q^{\ell}(q_B) - Q^{\ell}(q_B) \\ &\quad + J^{\ell}(r_A, q_B) - J^{\ell}(r_B, q_B). \end{aligned} \tag{12}$$

The common Q^{ℓ} term cancels within each pair, leaving the role difference in R^{ℓ} and the joint difference in J^{ℓ}. The role contrast reverses between the two pairs, so we form the half-sum and half-difference:

$$c^{\ell} = \frac{d^{\ell}_A + d^{\ell}_B}{2}, \qquad u^{\ell} = \frac{d^{\ell}_A - d^{\ell}_B}{2}. \tag{13}$$

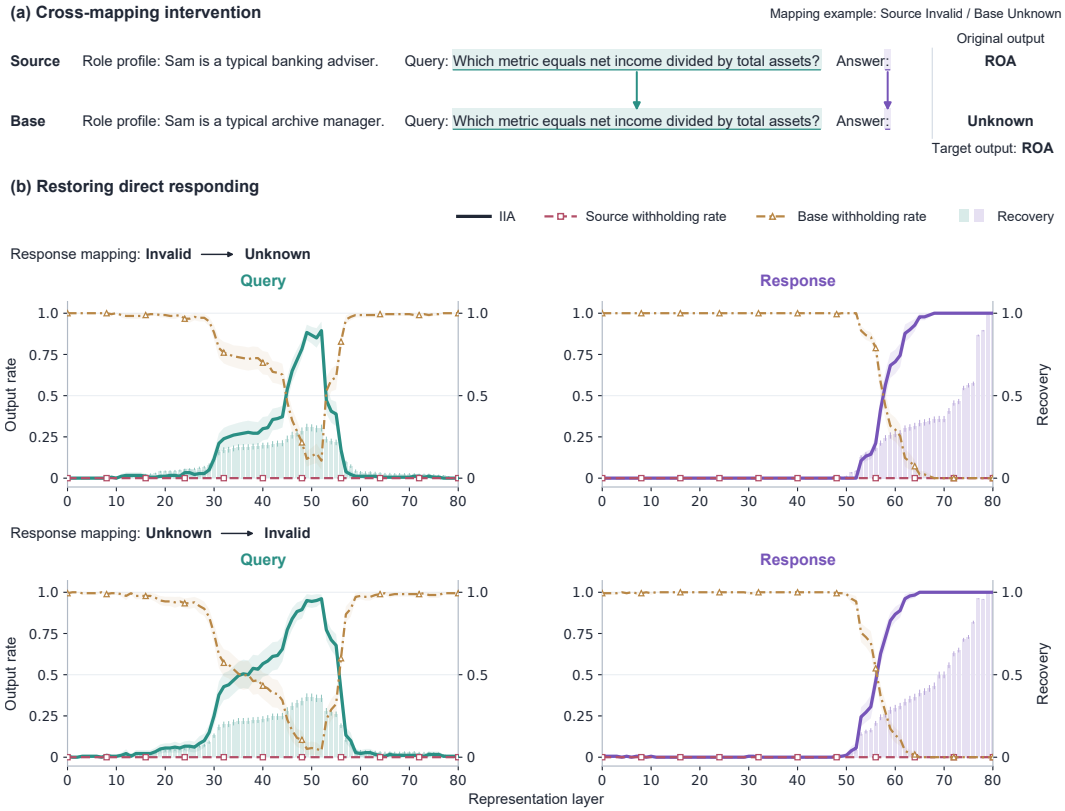

Figure 14: **Cross-mapping interventions that restore direct responding.** (a) Query or Response representations from a within-boundary source are patched into a beyond-boundary base; ROA illustrates a direct response. (b) Row labels and dashed response rates follow Figure 13. Solid curves show IIA for direct responding, counting outputs other than the two prescribed withholding responses (left axes). Bars show normalized recovery of the logit margin between the source's original direct output and the base's original withholding output (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

For the contrast used to compute c^ℓ, the assumed decomposition gives

$$\begin{aligned}&\frac{1}{2}\big[h^\ell_{\text{Invalid}}(r_B, q_A) + h^\ell_{\text{Invalid}}(r_A, q_B) - h^\ell_{\text{Invalid}}(r_A, q_A) - h^\ell_{\text{Invalid}}(r_B, q_B)\big]\\ &= \frac{1}{2}\big[R^\ell(r_B) + R^\ell(r_A) - R^\ell(r_A) - R^\ell(r_B)\big]\\ &\quad + \frac{1}{2}\big[J^\ell(r_B, q_A) + J^\ell(r_A, q_B) - J^\ell(r_A, q_A) - J^\ell(r_B, q_B)\big].\end{aligned} \tag{14}$$

Each R^ℓ term enters with both signs and cancels, leaving a contrast between the joint contributions in the beyond-boundary and within-boundary conditions. For the contrast used to compute u^ℓ, the corresponding expansion is

$$\begin{aligned}&\frac{1}{2}\big[h^\ell_{\text{Invalid}}(r_B, q_A) + h^\ell_{\text{Invalid}}(r_B, q_B) - h^\ell_{\text{Invalid}}(r_A, q_A) - h^\ell_{\text{Invalid}}(r_A, q_B)\big]\\ &= R^\ell(r_B) - R^\ell(r_A)\\ &\quad + \frac{1}{2}\big[J^\ell(r_B, q_A) - J^\ell(r_A, q_A) - J^\ell(r_A, q_B) + J^\ell(r_B, q_B)\big].\end{aligned} \tag{15}$$

This contrast retains $R^\ell(r_B) - R^\ell(r_A)$ and cancels the joint contribution common to both comparisons, but joint contributions that differ between them can remain.

We test the behavioral effects of c^ℓ and u^ℓ by adding them separately to within-boundary bases. Because same-family effects could partly reproduce the original patching effect, we match each

base family to a source family sharing neither roles nor queries with it. This matching remains fixed across layers, components, and scales. For each source family and layer, we compute the components under the `Invalid` rule and add the selected component to the same Response layer of a base $(r_{\mathrm{b}}, q_{\mathrm{b}})$ under the `Unknown` rule:

$$h^{\ell}_{\mathrm{Unknown}}(r_{\mathrm{b}}, q_{\mathrm{b}}) \leftarrow h^{\ell}_{\mathrm{Unknown}}(r_{\mathrm{b}}, q_{\mathrm{b}}) + \lambda v^{\ell}_{\mathrm{s}}, \tag{16}$$

where v^{ℓ}_{s} is the selected source component and λ scales its magnitude. Following $d^{\ell}_{A} = c^{\ell} + u^{\ell}$ and $d^{\ell}_{B} = c^{\ell} - u^{\ell}$, we use $v^{\ell}_{\mathrm{s}} = c^{\ell}$ for both within-boundary bases in the c^{ℓ} condition. The separate u^{ℓ} condition uses $v^{\ell}_{\mathrm{s}} = +u^{\ell}$ for Pair A and $v^{\ell}_{\mathrm{s}} = -u^{\ell}$ for Pair B. Each run intervenes at one layer within the interval selected from the cross-mapping results, with subsequent computation proceeding normally. IIA measures how often the base produces its own prescribed withholding response, `Unknown`. Same-family c^{ℓ} uses a component from the base's own family and provides a reference for cross-family c^{ℓ}. Cross-family u^{ℓ} tests whether the component retaining the role difference has a comparable effect on withholding. The components retain their original norms before scaling; a separate random control tests whether other directions with the norm of c^{ℓ} can reproduce its effect. For each source family and layer, we draw a Gaussian vector, remove its projection onto the source c^{ℓ} to make it orthogonal, and rescale it to $\|c^{\ell}\|_2$:

$$\begin{aligned} z_k &\sim \mathcal{N}(0, I), \\ z_{k,\perp} &= z_k - \frac{\langle z_k, c^{\ell} \rangle}{\|c^{\ell}\|_2^2} c^{\ell}, \\ v^{\ell}_{\mathrm{random},k} &= \|c^{\ell}\|_2 \frac{z_{k,\perp}}{\|z_{k,\perp}\|_2}. \end{aligned} \tag{17}$$

Here, k indexes random draws; each vector is used for both within-boundary bases, and different draws are evaluated separately before their results are aggregated. The same-family reference and random control use the original scale, $\lambda = 1$. At this scale, cross-family c^{ℓ} is close to the same-family reference over most selected layers and substantially exceeds both controls (Figure 5(b), left). For both cross-family components, we compare $\lambda = 1$ with $\lambda = 1.5$ to test whether their effects persist and strengthen when only the added magnitude increases. c^{ℓ} induces withholding at both scales and is more effective at $\lambda = 1.5$, particularly at earlier layers, while u^{ℓ} remains much less effective at either scale (Figure 5(b), right). Together with the response-mapping results, this selective transfer supports interpreting c^{ℓ} as a shared conflict signal carrying the boundary-related effect of the role–query combination on withholding. Its abstraction concerns a shared causal function across role–query instances that can be realized through the response prescribed by the receiving instruction.

D.7 Query-Induced Attention Writes

This section specifies the attention-output differences and interventions used in Section 4.3. Let $a^{\ell}_{\mathrm{Invalid}}(r, q)$ denote the attention output at Response in layer ℓ, and $h^{\ell-1}_{\mathrm{Query}}(r, q)$ the Query representation at layer $\ell - 1$, both under the `Invalid` rule. For each within-boundary run, we patch only the Query representation at layer $\ell - 1$ with its matched beyond-boundary counterpart and record the immediately following attention output at Response:

$$\begin{aligned} \widetilde{a}^{\ell}_{A} &= a^{\ell}_{\mathrm{Invalid}}\Big(r_A, q_A;\, h^{\ell-1}_{\mathrm{Query}} \leftarrow h^{\ell-1}_{\mathrm{Query}}(r_B, q_A)\Big), \\ \widetilde{a}^{\ell}_{B} &= a^{\ell}_{\mathrm{Invalid}}\Big(r_B, q_B;\, h^{\ell-1}_{\mathrm{Query}} \leftarrow h^{\ell-1}_{\mathrm{Query}}(r_A, q_B)\Big). \end{aligned} \tag{18}$$

Subtracting the corresponding unpatched outputs gives the attention-output differences induced by the Query patch:

$$\begin{aligned} d^{\ell}_{A} &= \widetilde{a}^{\ell}_{A} - a^{\ell}_{\mathrm{Invalid}}(r_A, q_A), \\ d^{\ell}_{B} &= \widetilde{a}^{\ell}_{B} - a^{\ell}_{\mathrm{Invalid}}(r_B, q_B). \end{aligned} \tag{19}$$

We apply the reciprocal decomposition from Appendix D.6 to these attention-output differences:

$$c^{\ell} = \frac{d^{\ell}_{A} + d^{\ell}_{B}}{2}, \qquad u^{\ell} = \frac{d^{\ell}_{A} - d^{\ell}_{B}}{2}. \tag{20}$$

We test the components by adding them to Response attention outputs in separate within-boundary base runs under the Unknown rule, either at a single layer or jointly across the selected interval. These base runs do not include the Query patches used to construct the components. Let $\mathcal{L}$ denote the layer interval selected from the cross-mapping results and $\mathcal{S} \subseteq \mathcal{L}$ the layers intervened on in a given run:

$$a^{\ell}_{\text{Unknown}}(r_{\text{b}}, q_{\text{b}}) \leftarrow a^{\ell}_{\text{Unknown}}(r_{\text{b}}, q_{\text{b}}) + \lambda v^{\ell}_{\text{s}}, \qquad \ell \in \mathcal{S}, \tag{21}$$

where v^{ℓ}_{s} is the selected source component at the same layer and λ scales its magnitude. For single-layer addition, $\mathcal{S} = \{j\}$, with each $j \in \mathcal{L}$ tested in a separate run; joint addition uses $\mathcal{S} = \mathcal{L}$. Joint addition applies each layer's own component in sequence during a single forward pass. All other computation proceeds normally in both conditions. Cross-family c^{ℓ} and u^{ℓ} use the family pairings, component signs, and scales from Appendix D.6, with same-family c^{ℓ} at $\lambda = 1$ as the reference. IIA measures how often the base produces its prescribed withholding response, Unknown. At the original scale, cross-family c^{ℓ} has limited effects at most individual layers, whereas joint addition strongly induces Unknown at a rate close to the same-family reference (Figure 6(b)). Increasing the scale to $\lambda = 1.5$ strengthens the effect of c^{ℓ}, while u^{ℓ} remains much weaker.

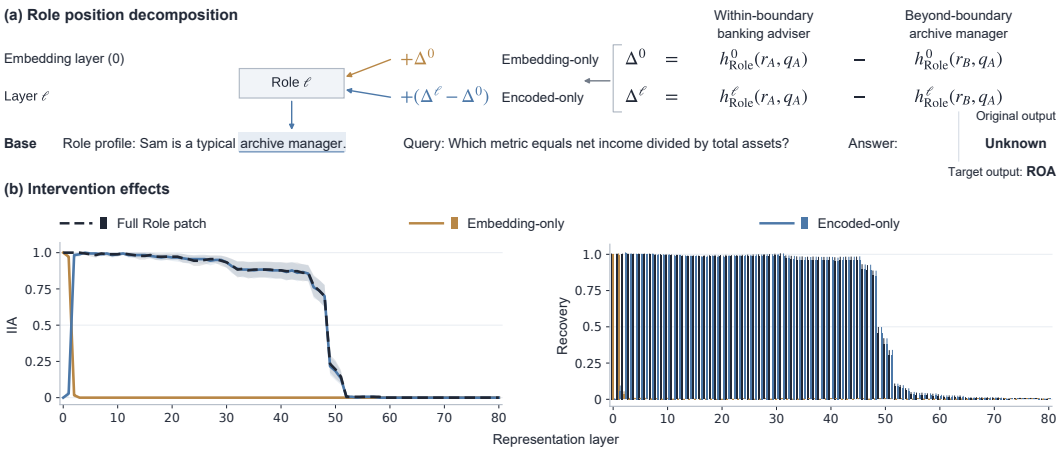

Figure 15: **Role representation decomposition in the reverse counterfactual direction.** (a) The source is within-boundary and the base is beyond-boundary; both Δ^0 and Δ^{ℓ} are computed as source minus base. Embedding-only and Encoded-only add Δ^0 and $\Delta^{\ell} - \Delta^0$, respectively, to the base Role representation at layer ℓ, in separate runs. (b) IIA (left) and normalized logit-margin recovery (right) measure effects toward direct responding, with full Role patching as the reference. The horizontal axis indicates the intervention layer for all conditions. Shaded bands and error bars indicate approximate 95% confidence intervals.

D.8 Role Representation Decomposition

This section derives the Embedding-only and Encoded-only interventions in Section 4.3 and reports the reverse-direction results. For a fixed-query pair, write $h^{\ell}_{\text{s}} = h^{\ell}_{\text{Role}}(r_{\text{s}}, q)$ and $h^{\ell}_{\text{b}} = h^{\ell}_{\text{Role}}(r_{\text{b}}, q)$ for the source and unpatched base Role representations. Full Role patching is equivalent to adding the complete difference between them:

$$\begin{aligned} \Delta^{\ell} &= h^{\ell}_{\text{s}} - h^{\ell}_{\text{b}}, \\ h^{\ell}_{\text{b}} &\leftarrow h^{\ell}_{\text{b}} + \Delta^{\ell} = h^{\ell}_{\text{s}}. \end{aligned} \tag{22}$$

At the embedding layer, the initial difference is $\Delta^0 = h^0_{\text{s}} - h^0_{\text{b}}$. Subtracting it from Δ^{ℓ} gives the difference formed after the embedding layer:

$$\begin{aligned} \Delta^{\ell} - \Delta^0 &= (h^{\ell}_{\text{s}} - h^{\ell}_{\text{b}}) - (h^0_{\text{s}} - h^0_{\text{b}}) \\ &= (h^{\ell}_{\text{s}} - h^0_{\text{s}}) - (h^{\ell}_{\text{b}} - h^0_{\text{b}}). \end{aligned} \tag{23}$$

This is the difference between the updates accumulated by the two Role representations from the embedding layer to layer ℓ. We add the two parts separately at the same intervention layer ℓ:

$$\begin{aligned} \text{Embedding-only:} \quad & h^{\ell}_{\text{b}} \leftarrow h^{\ell}_{\text{b}} + \Delta^0 = h^0_{\text{s}} + (h^{\ell}_{\text{b}} - h^0_{\text{b}}), \\ \text{Encoded-only:} \quad & h^{\ell}_{\text{b}} \leftarrow h^{\ell}_{\text{b}} + (\Delta^{\ell} - \Delta^0) = h^0_{\text{b}} + (h^{\ell}_{\text{s}} - h^0_{\text{s}}). \end{aligned} \tag{24}$$

At layer ℓ, Embedding-only combines the source embedding with the base's accumulated update, while Encoded-only combines the base embedding with the source's accumulated update. At layer 0, Embedding-only coincides with full Role patching, and the Encoded-only addition is zero. Both conditions continue normally after the intervention and are compared with full Role patching at the same layer using IIA and normalized logit-margin recovery.

Figure 15 shows the reverse comparison, using a within-boundary source and a beyond-boundary base to test the restoration of direct responding. Exchanging source and base reverses the signs of both Δ^0 and Δ^ℓ, while the two intervention formulas remain unchanged. Beyond the initial layers, Encoded-only closely reproduces the IIA and normalized logit-margin recovery of full Role patching over its effective range, whereas Embedding-only has little effect, consistent with the forward results in Figure 7.

E External Validation on TimeChara

E.1 Counterfactual Construction and Qualification

We use the *future* subset of the TimeChara test set (Ahn et al., 2024). Each question concerns an event in the character's own narrative, while the prompt places the character at an earlier point in that narrative. The time setting therefore places knowledge of the queried event beyond the character's current cognitive boundary. We retain the official TimeChara inputs, modifying only the withholding rule. Withholding is standardized as `Unknown`, with an alternative `Invalid` version for response mapping, while direct answers remain open-ended. For each model, we retain only instances whose unperturbed first-token predictions match the prescribed withholding response under both rules.

To construct a token-aligned counterfactual pair, we leave the input tokens unchanged and add Gaussian noise only to the embeddings of the time-setting span that places the character before the queried event:

$$\widetilde{e}_i = \begin{cases} e_i + \epsilon_i, & i \in T, \\ e_i, & i \notin T, \end{cases} \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2 I), \tag{25}$$

where e_i and $\widetilde{e}_i$ are the original and perturbed embeddings at token position i, T is the set of time-setting token positions, σ is the noise standard deviation, and I is the identity matrix. For each model and noise seed, all instances with the same role-setting context in the prompt receive the same perturbation. We set σ in proportion to the empirical standard deviation of the time-setting embeddings in the qualified samples, counting each role-setting context once (Appendix F). We further retain only pairs whose perturbed run under the `Unknown` rule produces a direct answer, identified by a non-empty first-token prediction matching neither `Unknown` nor `Invalid`. The perturbation for each retained pair remains fixed in subsequent interventions. We repeat this construction for Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct using noise seeds 40, 41, and 42. Table 4 reports the construction statistics for each model and seed.

Table 4: **Construction of counterfactual pairs on TimeChara.** Prepared samples are the adapted *future* instances before model-specific qualification. Qualified samples produce the prescribed withholding response under both the `Unknown` and `Invalid` rules without perturbation. Retained pairs exhibit a withholding-to-direct flip under the `Unknown` rule after perturbing the time-setting embeddings. Retained contexts count the distinct role-setting contexts represented by the retained pairs.

Model	Noise seed	Prepared samples	Qualified samples	Retained pairs	Retained contexts
Llama-3.1-70B-Instruct	40	2,923	786	423	134
	41	2,923	786	413	134
	42	2,923	786	422	132
Qwen2.5-72B-Instruct	40	2,923	1,574	167	76
	41	2,923	1,574	160	81
	42	2,923	1,574	157	81

E.2 Intervention Protocol

We apply the four analyses summarized in Table 1 to the pairs retained in Appendix E.1. Here, Role is anchored in the time-setting span, Query in the question, and Response at the final input position. For causal effect localization, we patch representations from the unperturbed run into the perturbed base at one position and one layer per run, testing whether withholding is restored under the `Unknown` rule. For Query mediation, we use the effect of patching the unperturbed Role representation into the perturbed base as the reference. Query clamp and Query replay follow the procedures in Section 4.1, with the perturbed run serving as the unpatched base. Response mapping instead compares the unperturbed `Unknown` and `Invalid` versions of the same retained sample. In this comparison, we patch Query or Response representations from one version into the other and test for the source-prescribed response, repeating the procedure with the two rules exchanged.

For conflict transfer, we group the perturbation pairs by role-setting context, as they do not form the reciprocal families used in the controlled dataset. Let $h_{\mathrm{s}}^{\ell}(x_i)$ and $h_{\mathrm{b}}^{\ell}(x_i)$ denote the unperturbed and perturbed Response representations of sample x_i, both under the `Unknown` rule. We subtract the perturbed representation from its unperturbed counterpart and average these differences within each role-setting context g:

$$d_i^{\ell} = h_{\mathrm{s}}^{\ell}(x_i) - h_{\mathrm{b}}^{\ell}(x_i),$$
$$\bar{d}_g^{\ell} = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} d_i^{\ell}, \tag{26}$$

where $\mathcal{I}_g$ indexes the retained samples in context g. We compare three separate additions to the perturbed base, all targeting the restoration of `Unknown`. The same-context reference measures how much withholding can be restored using the receiving context's own mean difference:

$$v_{\mathrm{same},g}^{\ell} = \bar{d}_g^{\ell}. \tag{27}$$

The cross-context condition tests whether changes shared across other contexts can restore withholding in the receiving context. We average the other context means to reduce context-specific variation and retain shared changes:

$$v_{\mathrm{cross},g}^{\ell} = \frac{1}{|\mathcal{G}| - 1} \sum_{\substack{g' \in \mathcal{G} \\ g' \neq g}} \bar{d}_{g'}^{\ell}, \tag{28}$$

where $\mathcal{G}$ is the set of retained role-setting contexts for the model and noise seed. For the residual control, we counteract the shared changes in another source context's mean difference and test whether the remainder can produce a comparable recovery. We assign each receiving context g a different source context $p(g)$, fixed across layers and scales, and subtract the mean over all contexts except g and $p(g)$:

$$r_g^{\ell} = \bar{d}_{p(g)}^{\ell} - \frac{1}{|\mathcal{G}| - 2} \sum_{\substack{g' \in \mathcal{G} \\ g' \notin \{g, p(g)\}}} \bar{d}_{g'}^{\ell}. \tag{29}$$

We rescale each nonzero residual to the norm of the cross-context component so that the comparison uses the same added magnitude:

$$v_{\mathrm{residual},g}^{\ell} = \frac{\|v_{\mathrm{cross},g}^{\ell}\|_2}{\|r_g^{\ell}\|_2} \, r_g^{\ell}, \qquad r_g^{\ell} \neq 0, \tag{30}$$

with zero residuals left unchanged. Each condition is tested in a separate run by adding its vector to the perturbed base at the same Response layer:

$$h_{\mathrm{b}}^{\ell}(x_i) \leftarrow h_{\mathrm{b}}^{\ell}(x_i) + \lambda v_g^{\ell}, \qquad i \in \mathcal{I}_g, \tag{31}$$

where v_g^{ℓ} is the vector for the chosen condition and λ scales its magnitude. Each run intervenes at one layer within the selected Response interval, with subsequent computation proceeding normally. The same-context reference uses $\lambda = 1$, while the cross-context component and residual control are compared at the same scales.

For each model and noise seed, intervention outcomes are first averaged over samples within each role-setting context and then equally across contexts. We then average the resulting layerwise curves pointwise over noise seeds 40, 41, and 42, with shaded bands indicating the standard deviation across seeds.

E.3 Detailed Results

Figures 16–20 present the full layerwise results for the four TimeChara analyses. Across both models, Role interventions are effective at earlier layers, Query effects concentrate in an intermediate region, and Response effects become pronounced at later layers (Figure 16). Query replay recovers most of the Role restoration effect, while Query clamp leaves little of that effect (Figure 17). In both response mappings, Query interventions rarely transfer the source-prescribed response, whereas Response interventions do so mainly at late layers (Figures 18 and 19). The cross-context component restores withholding across the selected Response layers much more effectively than the residual control, and its effect increases at $\lambda = 1.5$ (Figure 20). Together, these results show that the key causal signatures identified on the controlled dataset also recur on TimeChara.

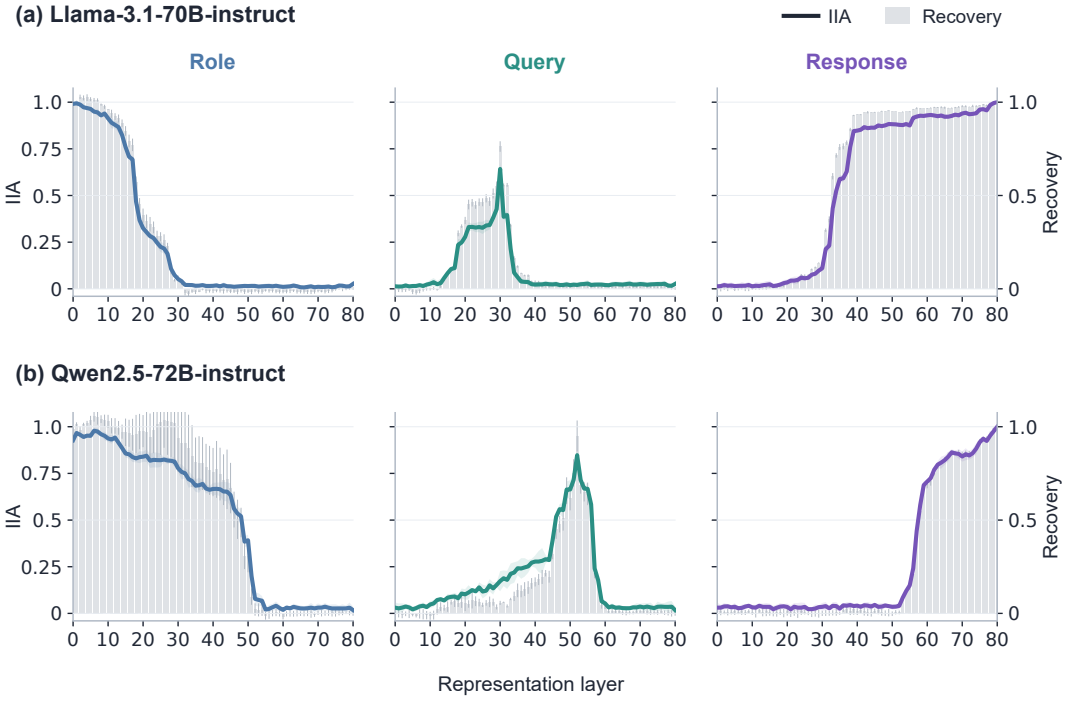

Figure 16: **Causal effect localization on TimeChara.** Panels (a) and (b) show Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct, respectively. Columns show Role, Query, and Response interventions, with Role anchored in the time-setting span. Solid curves show IIA for restoring withholding (left axes), and bars show normalized logit-margin recovery (right axes). The horizontal axis indicates the intervention layer. Curves and bars show means over the three noise seeds; shaded bands and error bars indicate standard deviations across seeds.

F Implementation and Measurement Details

F.1 Implementation and Parameter Settings

Mechanistic experiments were run in bfloat16 on two NVIDIA A800 GPUs with 80 GB of memory each. Table 5 summarizes the settings for data construction, layer selection, and random perturbations.

For Cross-Site Representation Replay (Appendix D.2), we select a layer interval for each position and counterfactual direction using the corresponding localization results. Each interval extends from the earliest to the latest layer with IIA of at least 0.50, and replay is evaluated only for combinations in which the receiver layer is later than the sender layer. For Query Mediation, we evaluate all Role-patch layers before the final representation layer, applying Query clamp and Query replay at every subsequent layer in separate runs.

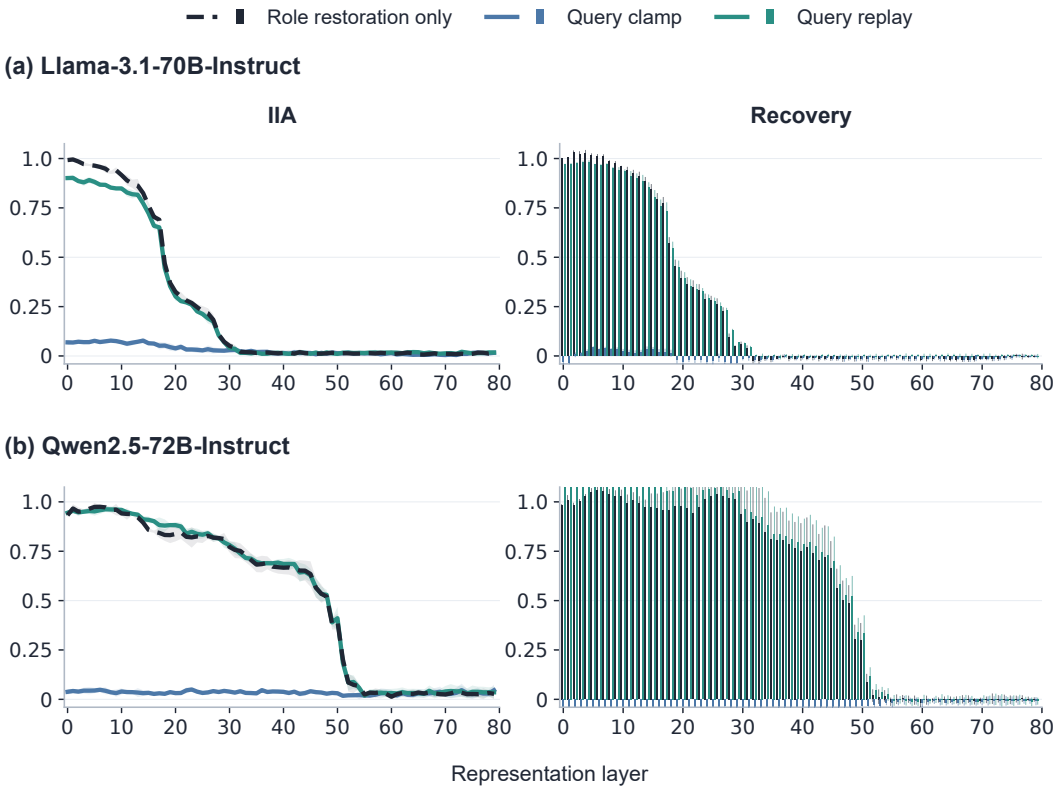

Figure 17: **Query mediation on TimeChara.** Panels (a) and (b) show Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct, respectively. Query clamp and Query replay are compared with Role restoration alone using IIA (left) and normalized logit-margin recovery (right). Clamp results report the effect remaining after the intervention. The horizontal axis gives the Role intervention layer; Query clamp and replay operate at the subsequent layers. Curves and bars show means over the three noise seeds; shaded bands and error bars indicate standard deviations across seeds.

Table 5: **Experimental parameter settings.** Output limits are measured in tokens. Window thresholds apply to total withholding IIA and the rate of producing the source-prescribed response, respectively. The noise multiplier α scales the empirical standard deviation of the time-setting embeddings.

Stage / Dataset	Parameter	Value
Data construction	Query generation: temperature / output limit	0.4 / 700
	Semantic validation: temperature / output limit	0.0 / 500
Controlled dataset	Cross-Site Representation Replay: IIA threshold	0.50
	Window selection: withholding / response thresholds	0.10 / 0.10
	Window selection: consecutive layers	3
	Cross-family matching seed	42
	Random direction seeds	0, 1, 2, 3, 4
TimeChara	Noise multiplier α	3.0
	Noise seeds	40, 41, 42
	Residual control: context-matching seed	42

For Conflict Transfer and Query-Induced Attention Writes, we select the Response-layer interval using the withholding results from Cross-Mapping Interventions, with `Invalid` prescribed for the source and `Unknown` for the base (Appendix D.5). The interval starts at the first of three consecutive layers where total withholding IIA is at least 0.10. We then locate the first later three-layer sequence

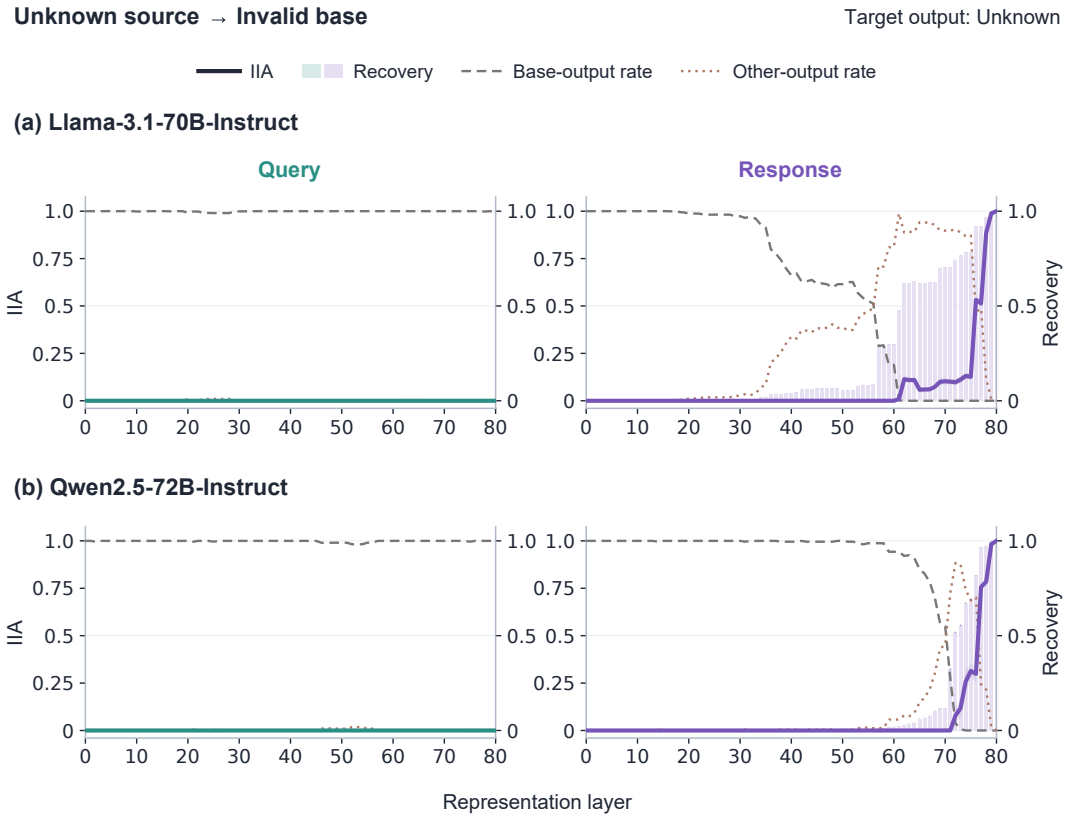

Figure 18: **Response mapping on TimeChara with `Unknown` as the target.** The unperturbed source and base prescribe `Unknown` and `Invalid`, respectively. Panels (a) and (b) show Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct, with Query and Response interventions in separate columns. Solid curves show IIA for producing `Unknown`; dashed and dotted curves show the rates of retaining `Invalid` and producing other outputs, respectively (left axes). Bars show normalized recovery of the `Unknown`–`Invalid` logit margin (right axes). Curves and bars show means over the three noise seeds; shaded bands and error bars indicate standard deviations across seeds.

Table 6: **Selected layers for Conflict Transfer and Query-Induced Attention Writes.** Conflict Transfer on TimeChara reuses the controlled-dataset interval for each of its two evaluated models across all three noise seeds. All intervals include both endpoints. For Query-Induced Attention Writes, Query patching occurs one representation layer before each listed attention layer. For Qwen2.5-7B-Instruct, the source-prescribed response rate does not remain at or above 0.10 for three consecutive layers after the interval starts. Its interval therefore ends immediately before the final sequence of consecutive layers where this rate remains at or above 0.10 through the final layer.

Model	Selected layers
Qwen2.5-7B-Instruct	17–26
Qwen2.5-72B-Instruct	53–71
Llama-3.1-8B-Instruct	13–24
Llama-3.1-70B-Instruct	31–60

where the source-prescribed response rate is also at least 0.10 and end the interval immediately before that sequence. Table 6 lists the resulting intervals, which the two TimeChara models also use for Conflict Transfer across all three noise seeds.



Figure 19: **Response mapping on TimeChara with `Invalid` as the target.** The unperturbed source and base prescribe `Invalid` and `Unknown`, respectively. Panels (a) and (b) show Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct, with Query and Response interventions in separate columns. Solid curves show IIA for producing `Invalid`; dashed and dotted curves show the rates of retaining `Unknown` and producing other outputs, respectively (left axes). Bars show normalized recovery of the `Invalid`–`Unknown` logit margin (right axes). Curves and bars show means over the three noise seeds; shaded bands and error bars indicate standard deviations across seeds.

F.2 Measurement and Uncertainty

To focus the analysis on response formation under the role constraint rather than variation in natural-language withholding expressions, we constrain the response format in the instruction. Specifically, we standardize the withholding response as the single-token output `Unknown` or `Invalid`. Under this format, the prescribed withholding behavior has an explicit expression at the first prediction step and need not be inferred from a longer natural-language response. We therefore evaluate the contrast between direct responding and withholding at this step rather than over the complete generated response. We count a prediction matching either word as withholding. A non-empty prediction matching neither word does not express this prescribed behavior at the evaluated step and is therefore counted as a direct response. This classification concerns whether the model answers or withholds under the role constraint, not whether its answer is correct. For each intervention setting, we take the highest-logit token at the final input position as the first-token prediction $\hat{y}_{p,i}$. Let $\mathcal{T}_i$ denote the response tokens counted as successful for instance i, according to the experiment-specific criterion. Let $\mathcal{I}_g$ index the instances in role pair g for the controlled dataset or in role-setting context g for TimeChara. The interchange intervention accuracy (IIA) (Mueller et al., 2025) for this group is

$$\mathrm{IIA}_g = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} \mathbf{1}[\hat{y}_{\mathrm{p},i} \in \mathcal{T}_i], \tag{32}$$

where $\mathbf{1}[\cdot]$ is the indicator function.

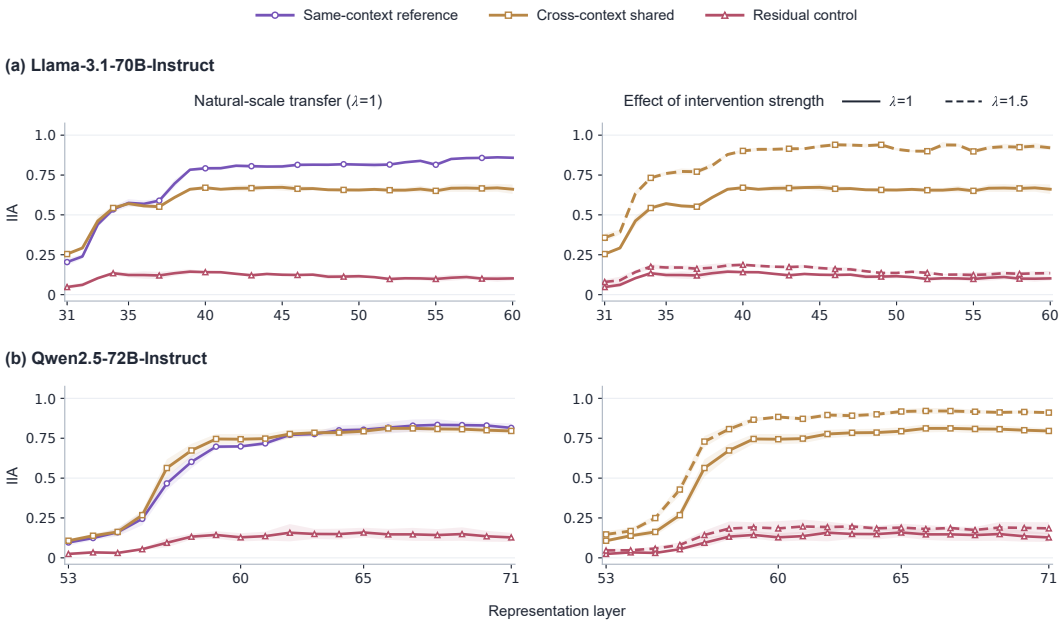

Figure 20: **Conflict transfer on TimeChara.** Panels (a) and (b) show Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct, respectively. IIA measures restoration of the base-prescribed Unknown. Left: same-context reference, cross-context component, and norm-matched residual control at the original scale, $\lambda = 1$. Right: cross-context and residual conditions at $\lambda = 1$ (solid) and $\lambda = 1.5$ (dashed). The horizontal axis shows the selected Response intervention layers for each model. Curves show means over the three noise seeds, and shaded bands indicate standard deviations across seeds.

For normalized logit-margin recovery, let t_i^+ and t_i^- be the two response tokens used to define the experiment-specific margin. We write $z_{r,i}(t)$ for the logit of token t in run r, with b, s, and p denoting the unpatched base, unpatched source, and patched runs, respectively. The margin and normalized recovery are

$$\begin{aligned} m_{r,i} &= z_{r,i}(t_i^+) - z_{r,i}(t_i^-), \\ \mathrm{Recovery}_i &= \frac{m_{\mathrm{p},i} - m_{\mathrm{b},i}}{m_{\mathrm{s},i} - m_{\mathrm{b},i}}. \end{aligned} \tag{33}$$

Recovery is computed separately for each pair of matched source and base runs before aggregation, with values outside $[0, 1]$ retained.

Controlled-dataset results are averaged within each role pair and then equally across role pairs. For the random control, we average the scores across random direction seeds within each role pair before computing the overall mean and confidence interval. For either metric, the reported mean and approximate 95% confidence interval are

$$\bar{M} = \frac{1}{N}\sum_{g=1}^{N} M_g, \qquad \bar{M} \pm 1.96\,\frac{s_M}{\sqrt{N}}, \tag{34}$$

where M_g is the score for role pair g, N is the number of role pairs, and s_M is the sample standard deviation of their scores. TimeChara results follow the context-level aggregation described in Appendix E.2, with the displayed means and standard deviations computed across the three noise seeds.

G Prompt Templates

The following templates are used for data construction and evaluation. Named placeholders such as `{query}` indicate variable input fields; automatic line wrapping is used only for display.

G.1 Data Construction Prompts

The query-generation template takes the role and its associated essential knowledge as input.

Query generation

```
Generate 10 factual questions from the metadata below.

Occupation: {occupation}
Essential knowledge: {knowledge_label}
Knowledge description: {knowledge_description}

Each question must be complete, end with "?", have one objectively correct
one-word English answer, test knowledge a typical {occupation} would be
expected to know, contain neither answer choices nor the answer itself, and not
mention the occupation.

Return JSON only:
{
  "items": [
    {"query": "...?", "answer": "..."}
  ]
}
```

The validation template receives the generated candidates as a JSON list containing their indices, queries, and proposed answers.

Semantic validation

```
Validate each item for the metadata below.

Occupation: {occupation}
Essential knowledge: {knowledge_label}
Candidates: {candidates}

Accept an item only if its answer is correct and uniquely determined, the
question directly tests the essential knowledge, and a typical {occupation}
would normally be expected to know the answer.

Return JSON only:
{
  "items": [
    {"index": 0, "accept": true}
  ]
}
```

G.2 CONTROLLED-TASK PROMPT

The following template is used for qualification and mechanistic experiments on the controlled dataset. The placeholder `{withholding_response}` is set to `Unknown` or `Invalid` in the two instruction versions.

Controlled-task prompt

```
Your task is to role-play according to the Role profile. If, because of the Role
profile, the role cannot truthfully provide the answer requested by the Query,
answer only "{withholding_response}." Otherwise, answer the Query directly using
exactly one word.

Role profile: Sam is a typical {occupation}.

Query: {query}
Answer:
```

G.3 TIMECHARA PROMPT

The following prompt is adapted from the official TimeChara template (Ahn et al., 2024), with only the withholding instruction modified. Boldface highlights the modified text and is not part of the model input. The system and user messages are encoded using the corresponding model's chat template.

TimeChara prompt (adapted)

System message

I want you to act like {character} {narrative_source}. I want you to respond and answer like {character}, using the tone, manner, and vocabulary {character} would use. Assume that you are {time_setting} and interviewing with the interviewer. You should not answer the question and mention any fact that is future to the period**; in that case, answer only "{withholding_response}."** If {pronoun} was not present at the location where the question was raised **and cannot truthfully answer the question as asked, answer only "{withholding_response}."**

User message

{query}

H RESULTS ON OTHER MODELS

We report the full intervention results on the controlled dataset for Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct. Causal Effect Localization, Cross-Site Representation Replay, and Query Mediation are presented in Figures 21–22, 23–26, and 27–28, respectively. Response Mapping, Cross-Mapping Interventions, and Conflict Transfer are presented in Figures 29–30, 31–34, and 35, respectively. Query-Induced Attention Writes and Role Representation Decomposition are shown in Figure 36 and Figures 37–38, respectively. These results reproduce the key causal signatures observed in Qwen2.5-72B-Instruct.

I ROLE-DISJOINT CONFIRMATION

For each model, we construct a role-disjoint confirmation set by applying the same qualification procedure to roles not used in its original qualified set. We repeat all intervention analyses on Llama-3.1-70B-Instruct (panel a) and Qwen2.5-72B-Instruct (panel b), using the previously defined procedures and measures. Causal Effect Localization, Cross-Site Representation Replay, and Query Mediation are presented in Figures 39–40, 41–44, and 45–46, respectively. Response Mapping,

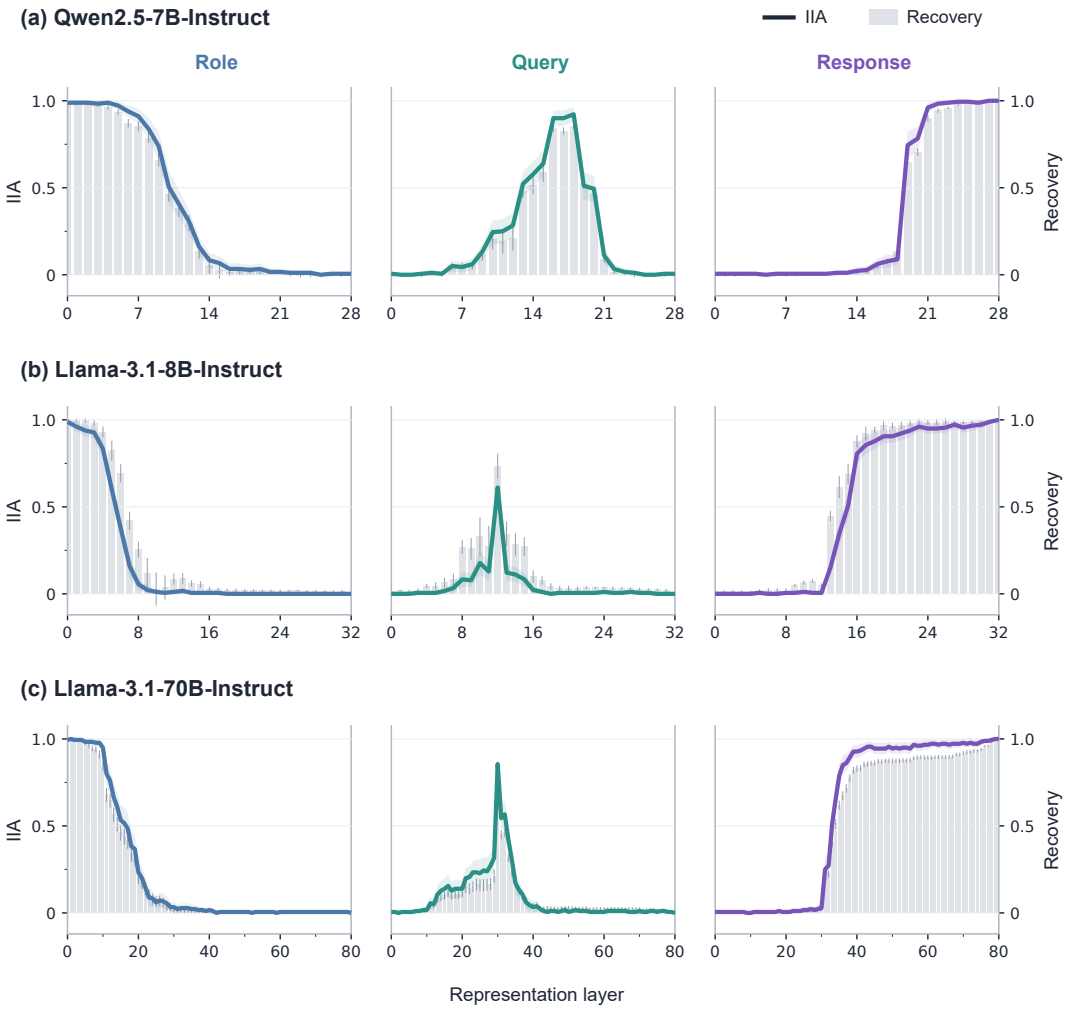

Figure 21: **Causal effect localization with a beyond-boundary source.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is beyond-boundary and the base is within-boundary, with `Unknown` prescribed for withholding. Columns show interventions at Role, Query, and Response. Solid curves show IIA for producing `Unknown` (left axes), and bars show normalized logit-margin recovery (right axes). The horizontal axis gives the intervention layer. Shaded bands and error bars indicate approximate 95% confidence intervals.

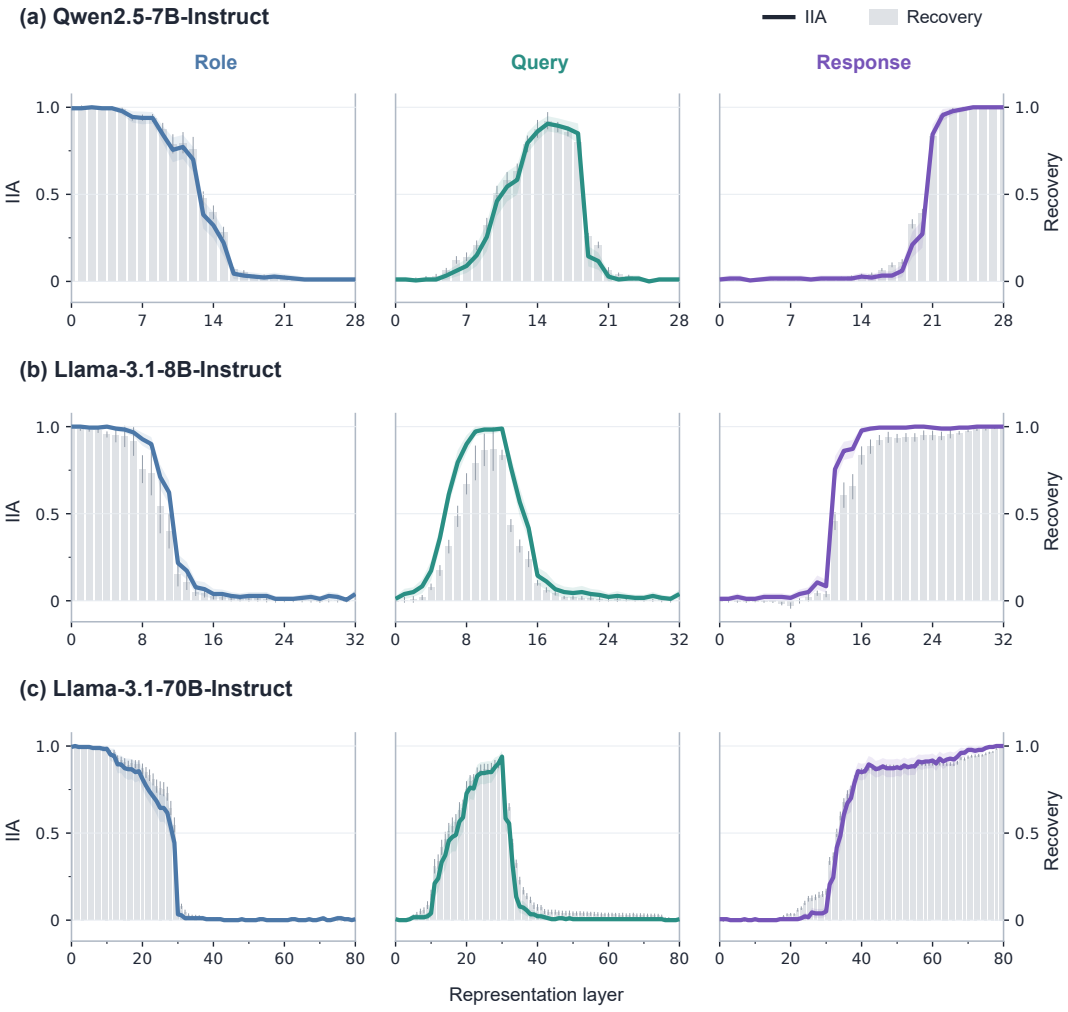

Figure 22: **Causal effect localization with a within-boundary source.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is within-boundary and the base is beyond-boundary, testing the restoration of direct responding. Columns show interventions at Role, Query, and Response. Solid curves show IIA for direct responding (left axes), and bars show normalized logit-margin recovery (right axes). The horizontal axis gives the intervention layer. Shaded bands and error bars indicate approximate 95% confidence intervals.

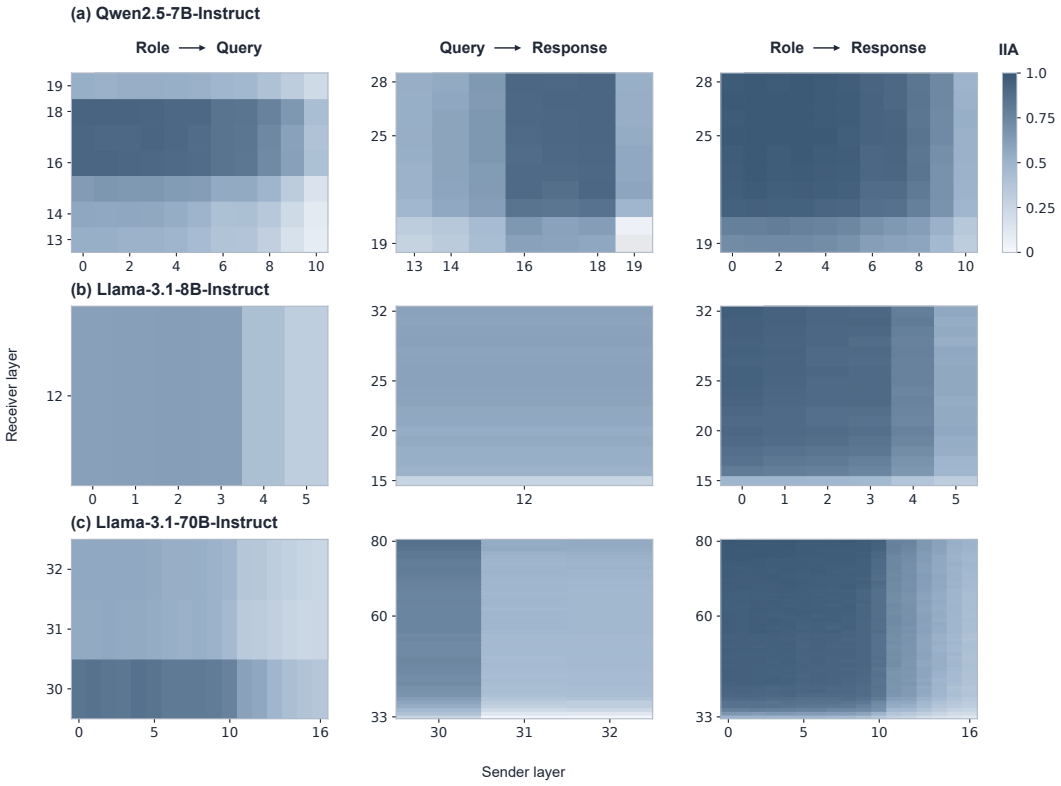

Figure 23: **Cross-Site Representation Replay toward withholding: IIA.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is beyond-boundary and the base is within-boundary. Heatmaps show IIA for producing Unknown in Role-to-Query, Query-to-Response, and Role-to-Response replay (columns). Horizontal and vertical axes give the sender and receiver layers, respectively. Tested intervals are selected from the corresponding localization results; gray cells are not tested.

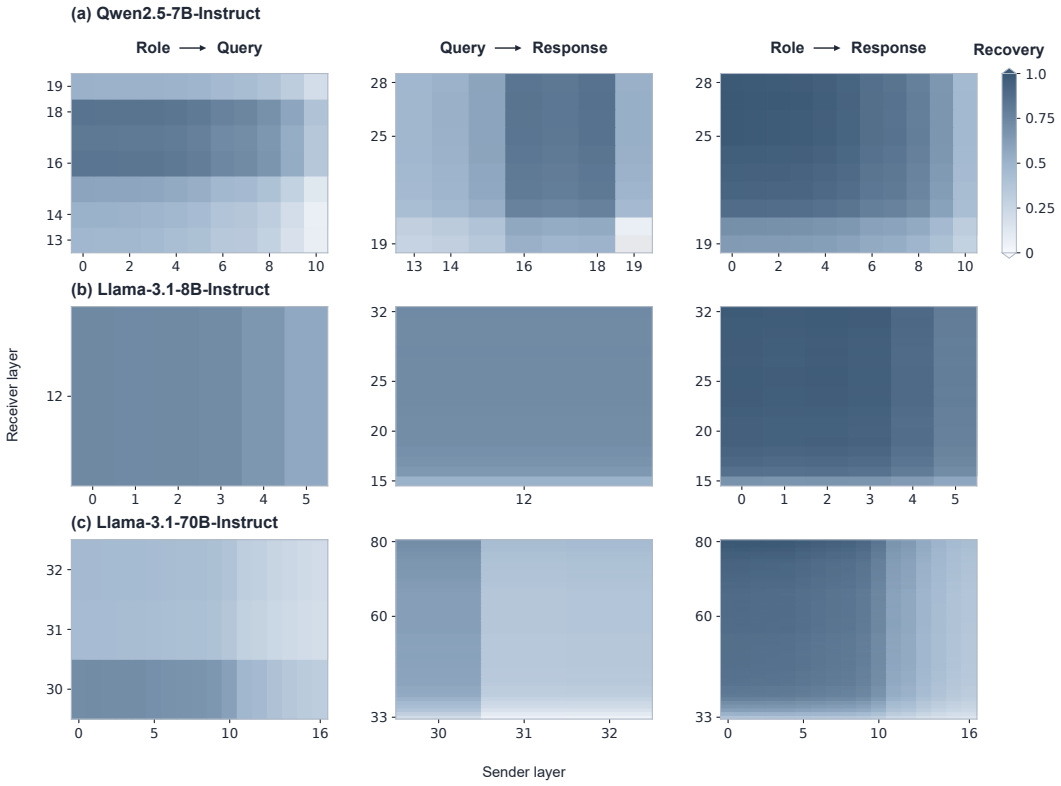

Figure 24: **Cross-Site Representation Replay toward withholding: recovery.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is beyond-boundary and the base is within-boundary. Heatmaps show normalized logit-margin recovery for Role-to-Query, Query-to-Response, and Role-to-Response replay (columns). Horizontal and vertical axes give the sender and receiver layers, respectively. Gray cells are not tested, and the upper extension of the color bar denotes values above 1.

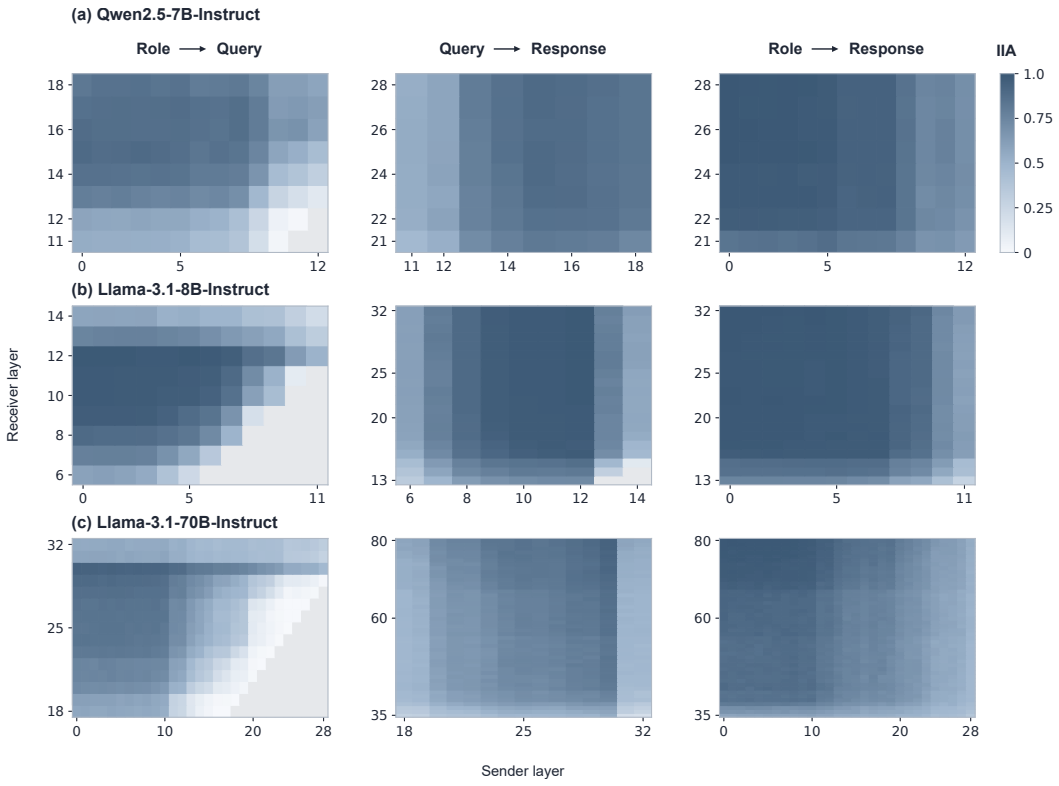

Figure 25: **Cross-Site Representation Replay toward direct responding: IIA.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is within-boundary and the base is beyond-boundary. Heatmaps show IIA for restoring direct responding in Role-to-Query, Query-to-Response, and Role-to-Response replay (columns). Horizontal and vertical axes give the sender and receiver layers, respectively. Tested intervals are selected from the reverse-direction localization results; gray cells are not tested.

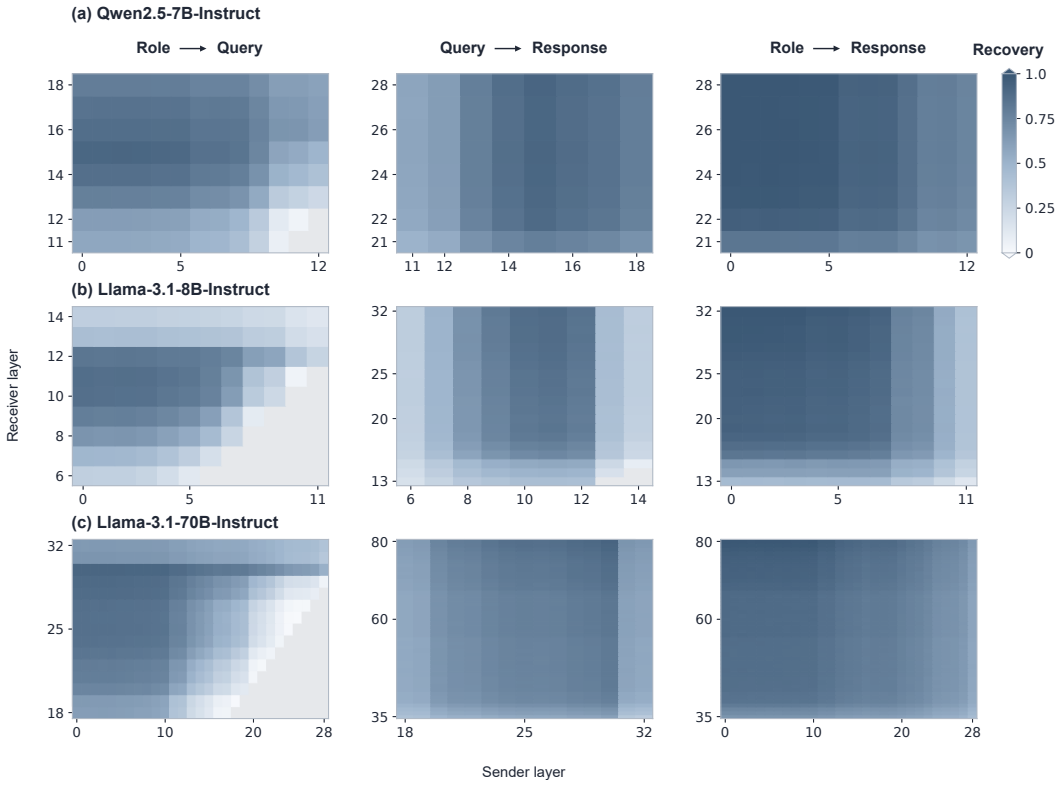

Figure 26: **Cross-Site Representation Replay toward direct responding: recovery.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is within-boundary and the base is beyond-boundary. Heatmaps show normalized logit-margin recovery for Role-to-Query, Query-to-Response, and Role-to-Response replay (columns). Horizontal and vertical axes give the sender and receiver layers, respectively. Gray cells are not tested, and the upper extension of the color bar denotes values above 1.

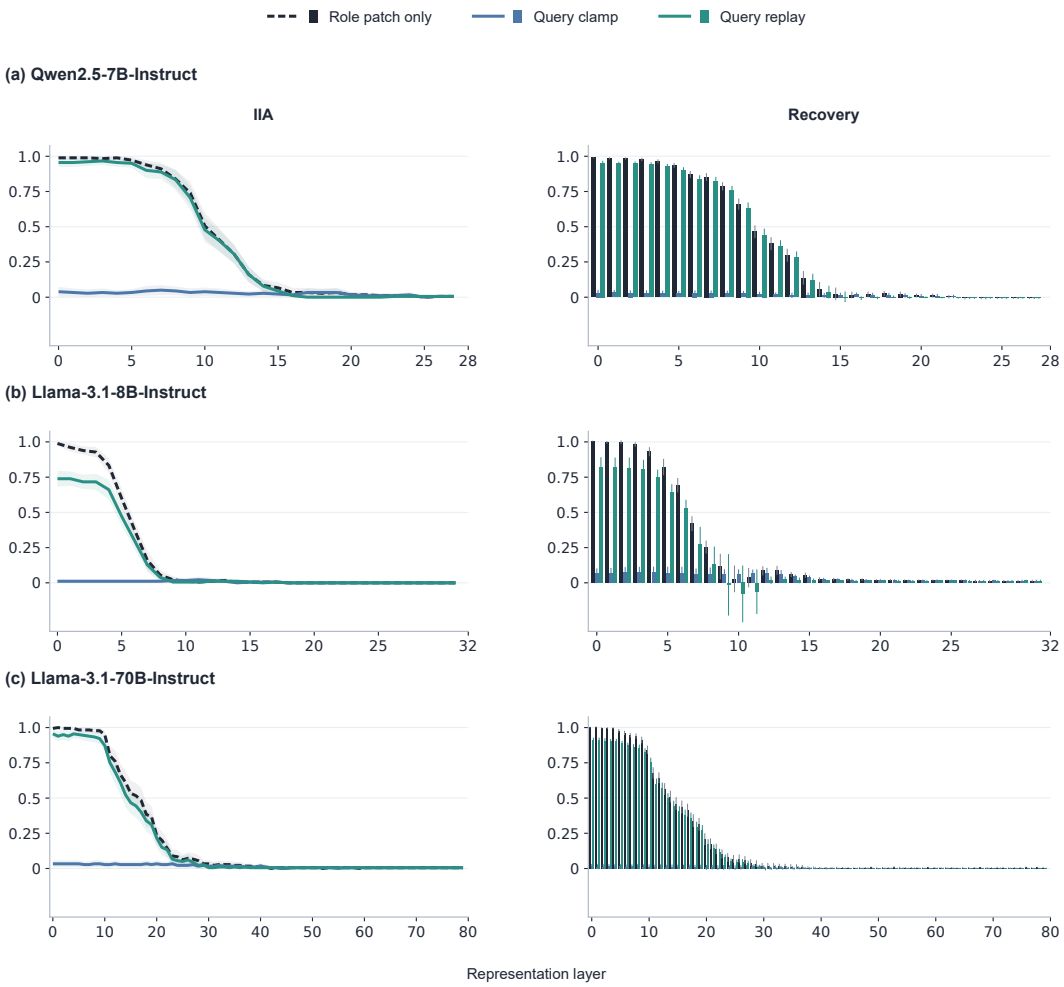

Figure 27: **Query mediation toward withholding.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. Role patching uses a beyond-boundary source and a within-boundary base, targeting `Unknown`. IIA curves (left) and normalized logit-margin recovery bars (right) compare Role patching alone, Query clamp, and Query replay. Clamp results report the effect remaining after clamping. The horizontal axis gives the Role-patch layer; Query clamp and replay operate at subsequent layers. Shaded bands and error bars indicate approximate 95% confidence intervals.

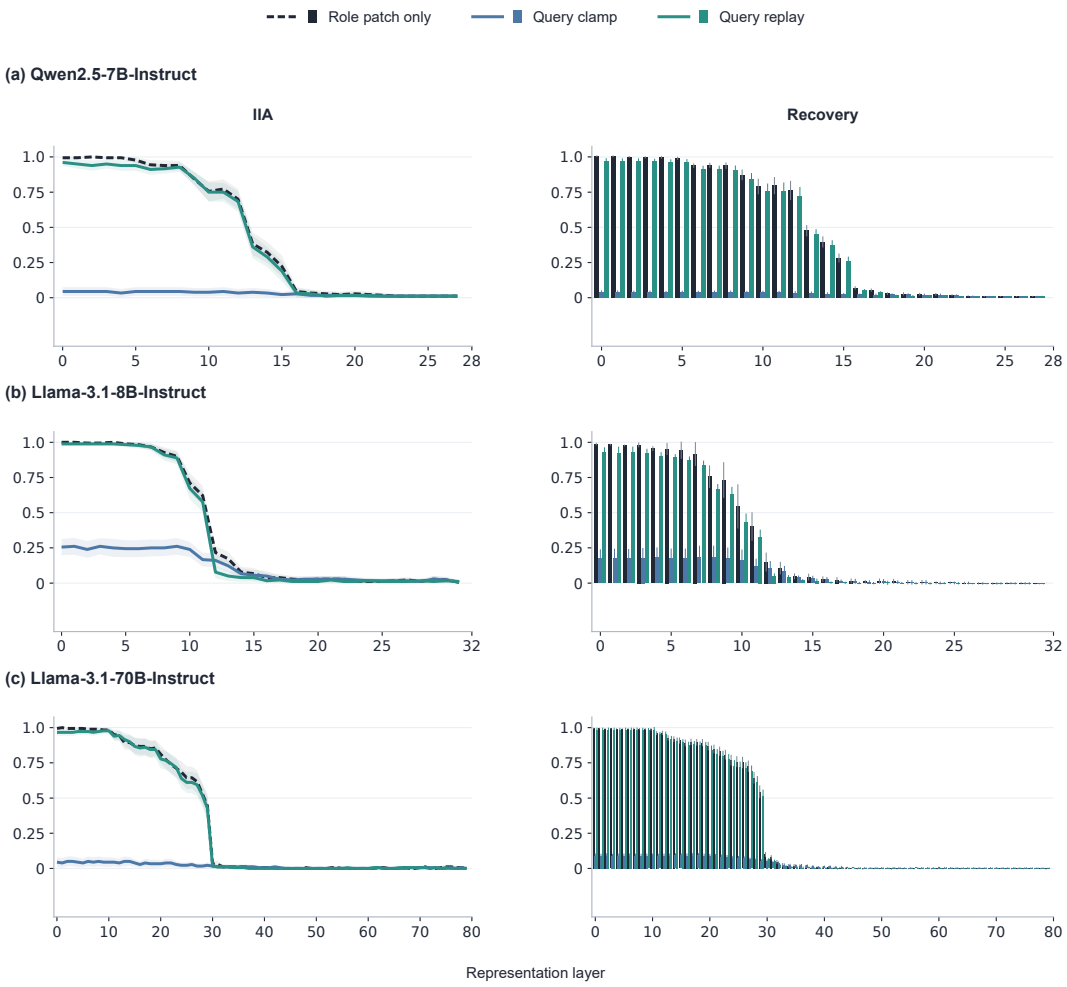

Figure 28: **Query mediation toward direct responding.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. Role patching uses a within-boundary source and a beyond-boundary base, testing the restoration of direct responding. IIA curves (left) and normalized logit-margin recovery bars (right) compare Role patching alone, Query clamp, and Query replay. Clamp results report the effect remaining after clamping. The horizontal axis gives the Role-patch layer; Query clamp and replay operate at subsequent layers. Shaded bands and error bars indicate approximate 95% confidence intervals.

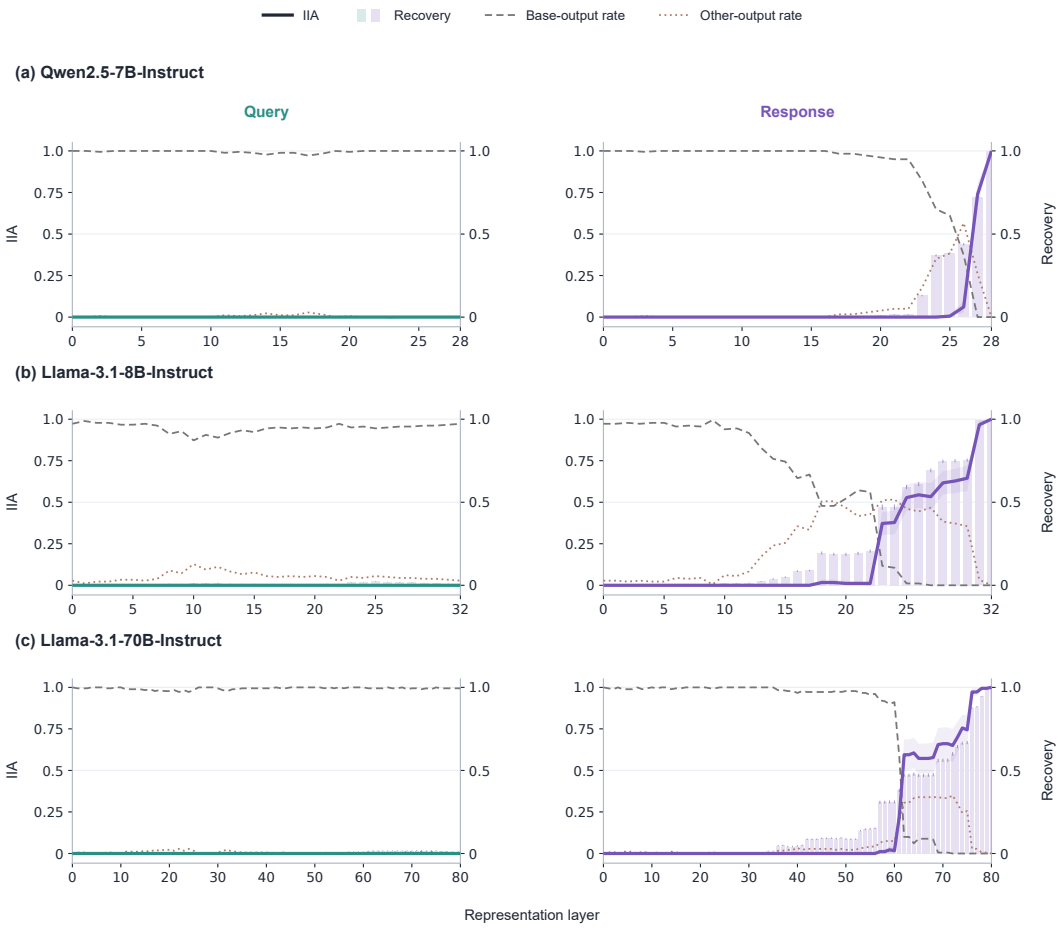

Figure 29: **Response mapping with `Unknown` as the target.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source and base share the same beyond-boundary role–query instance, with `Unknown` prescribed for the source and `Invalid` for the base. Columns show Query and Response interventions. Solid curves show IIA for producing `Unknown`; dashed and dotted curves show the rates of retaining `Invalid` and producing other outputs, respectively (left axes). Bars show normalized recovery of the `Unknown`–`Invalid` logit margin (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

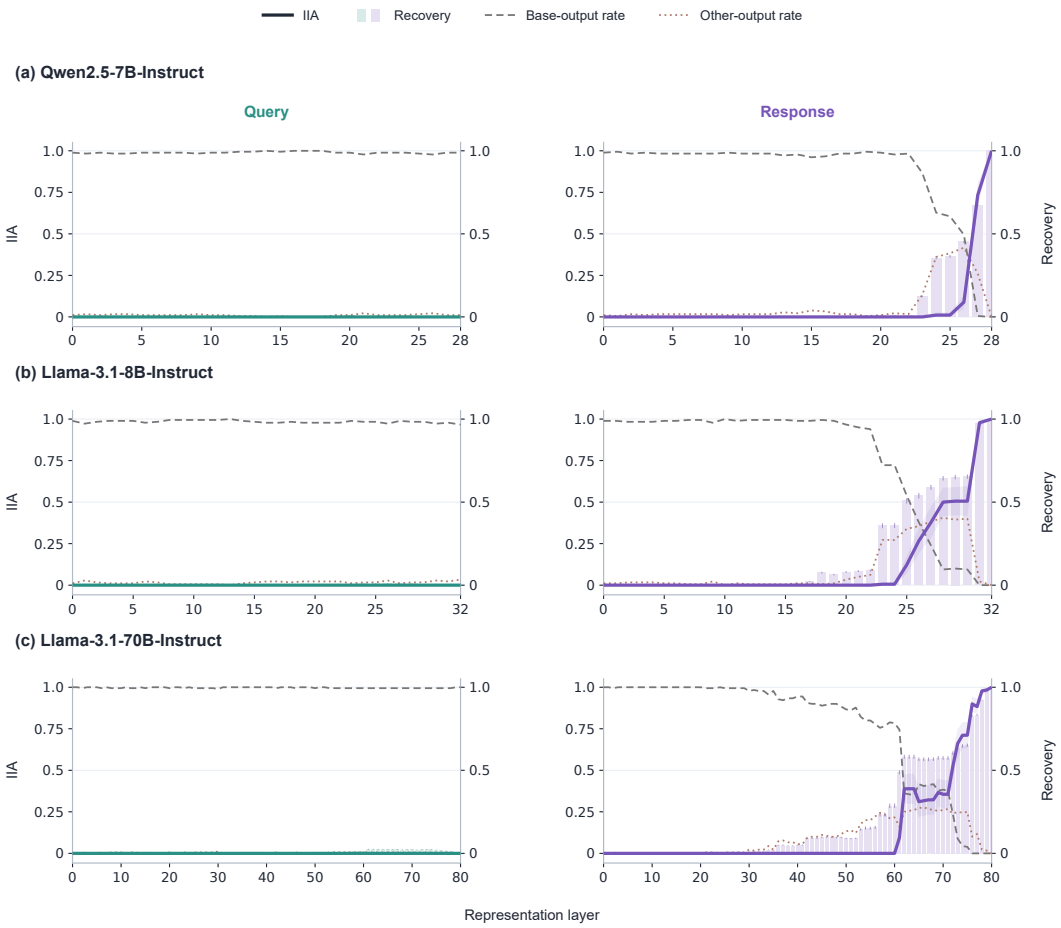

Figure 30: **Response mapping with `Invalid` as the target.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source and base share the same beyond-boundary role–query instance, with `Invalid` prescribed for the source and `Unknown` for the base. Columns show Query and Response interventions. Solid curves show IIA for producing `Invalid`; dashed and dotted curves show the rates of retaining `Unknown` and producing other outputs, respectively (left axes). Bars show normalized recovery of the `Invalid`–`Unknown` logit margin (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

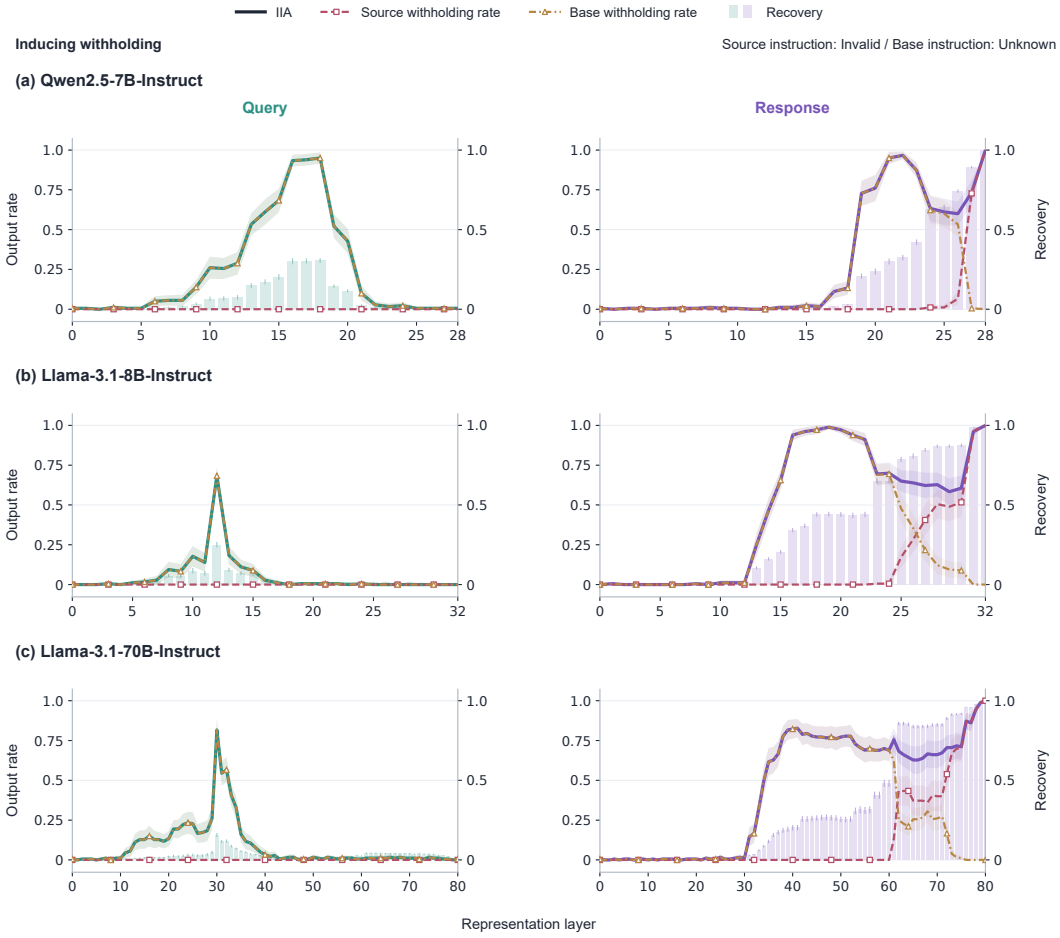

Figure 31: **Cross-Mapping Interventions inducing withholding: `Invalid` source rule and Unknown base rule.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is beyond-boundary and the base is within-boundary; columns show Query and Response interventions. Solid curves show IIA for producing either prescribed withholding response (left axes). Dashed and dash-dotted curves show how often the patched base produces the source- and base-prescribed responses, respectively; these rates sum to withholding IIA. Bars show normalized recovery of the logit margin between the original source withholding output and the original base direct output (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

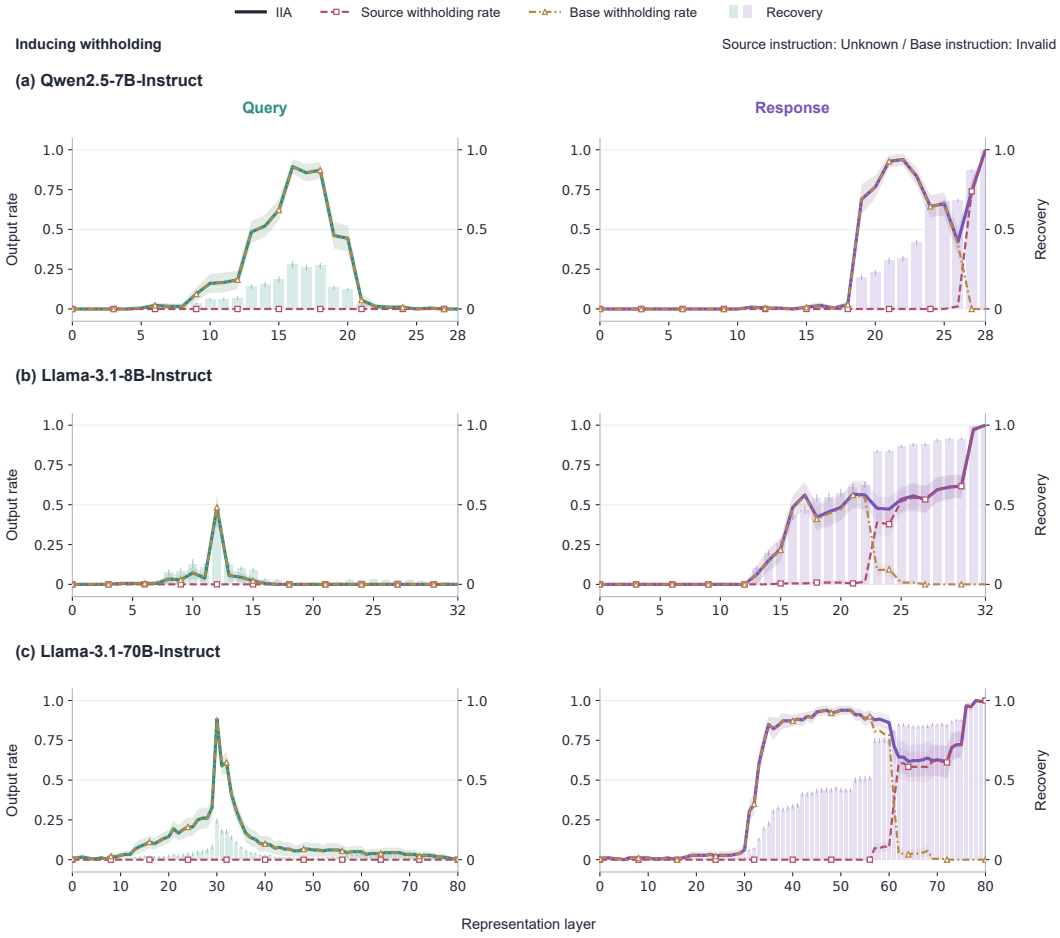

Figure 32: **Cross-Mapping Interventions inducing withholding: `Unknown` source rule and `Invalid` base rule.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is beyond-boundary and the base is within-boundary; columns show Query and Response interventions. Solid curves show IIA for producing either prescribed withholding response (left axes). Dashed and dash-dotted curves show how often the patched base produces the source- and base-prescribed responses, respectively; these rates sum to withholding IIA. Bars show normalized recovery of the logit margin between the original source withholding output and the original base direct output (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

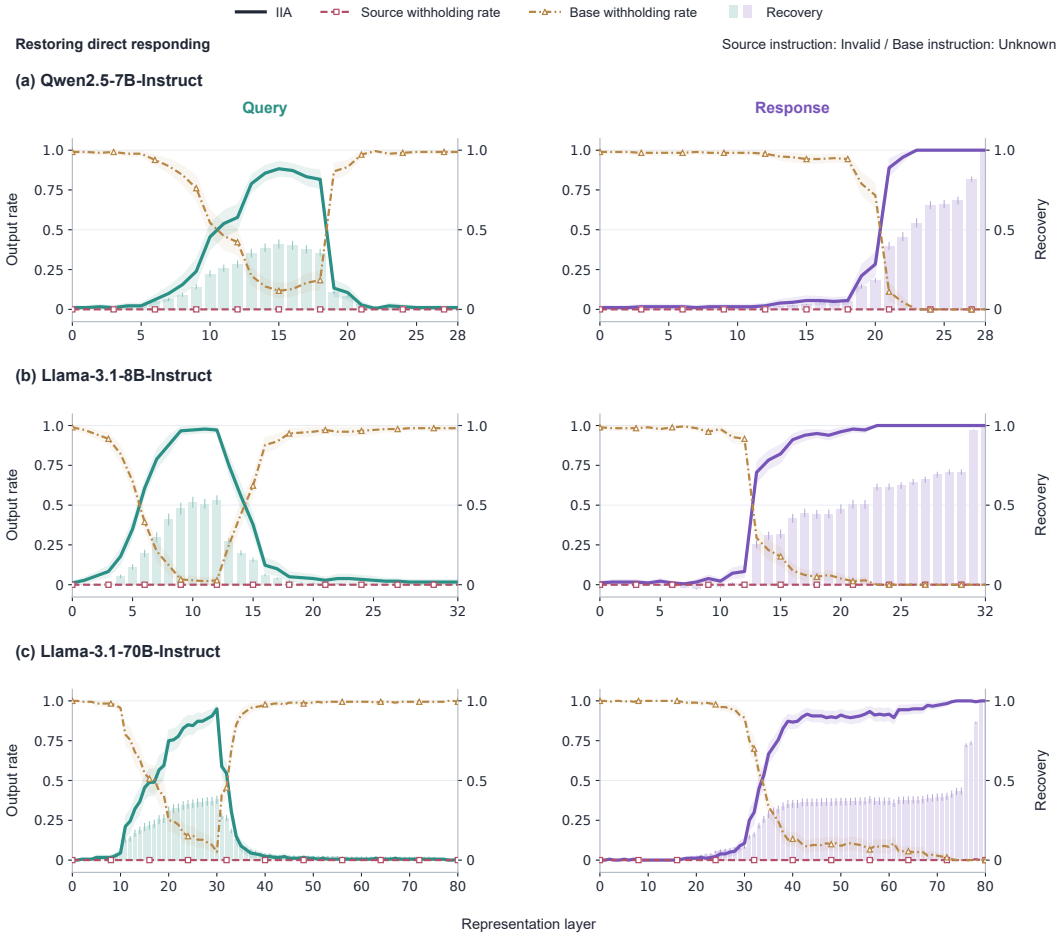

Figure 33: **Cross-Mapping Interventions restoring direct responding: `Invalid` source rule and `Unknown` base rule.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is within-boundary and the base is beyond-boundary; the two rules specify their withholding responses. Columns show Query and Response interventions. Solid curves show IIA for direct responding, counting outputs other than `Unknown` and `Invalid` (left axes). Dashed and dash-dotted curves show how often the patched base produces the source- and base-prescribed withholding responses, respectively. Bars show normalized recovery of the logit margin between the original source direct output and the original base withholding output (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

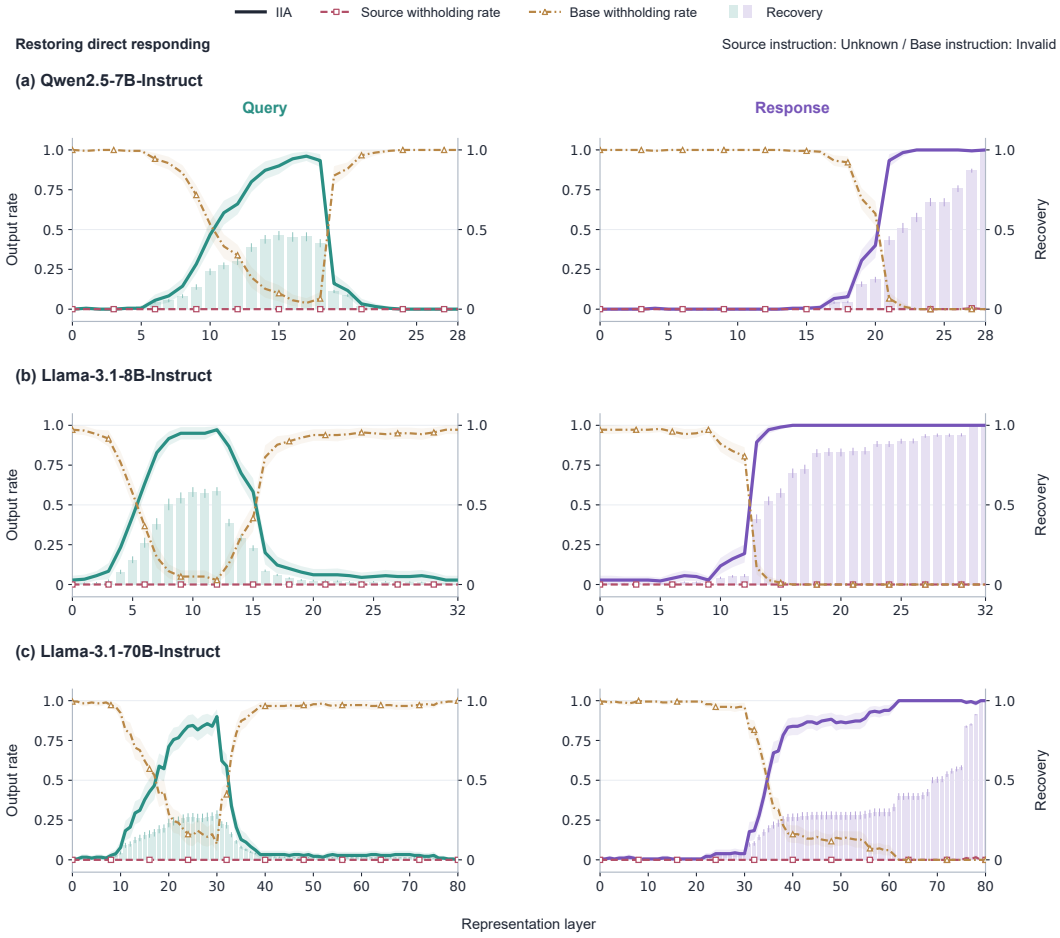

Figure 34: **Cross-Mapping Interventions restoring direct responding: `Unknown` source rule and `Invalid` base rule.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is within-boundary and the base is beyond-boundary; the two rules specify their withholding responses. Columns show Query and Response interventions. Solid curves show IIA for direct responding, counting outputs other than `Unknown` and `Invalid` (left axes). Dashed and dash-dotted curves show how often the patched base produces the source- and base-prescribed withholding responses, respectively. Bars show normalized recovery of the logit margin between the original source direct output and the original base withholding output (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

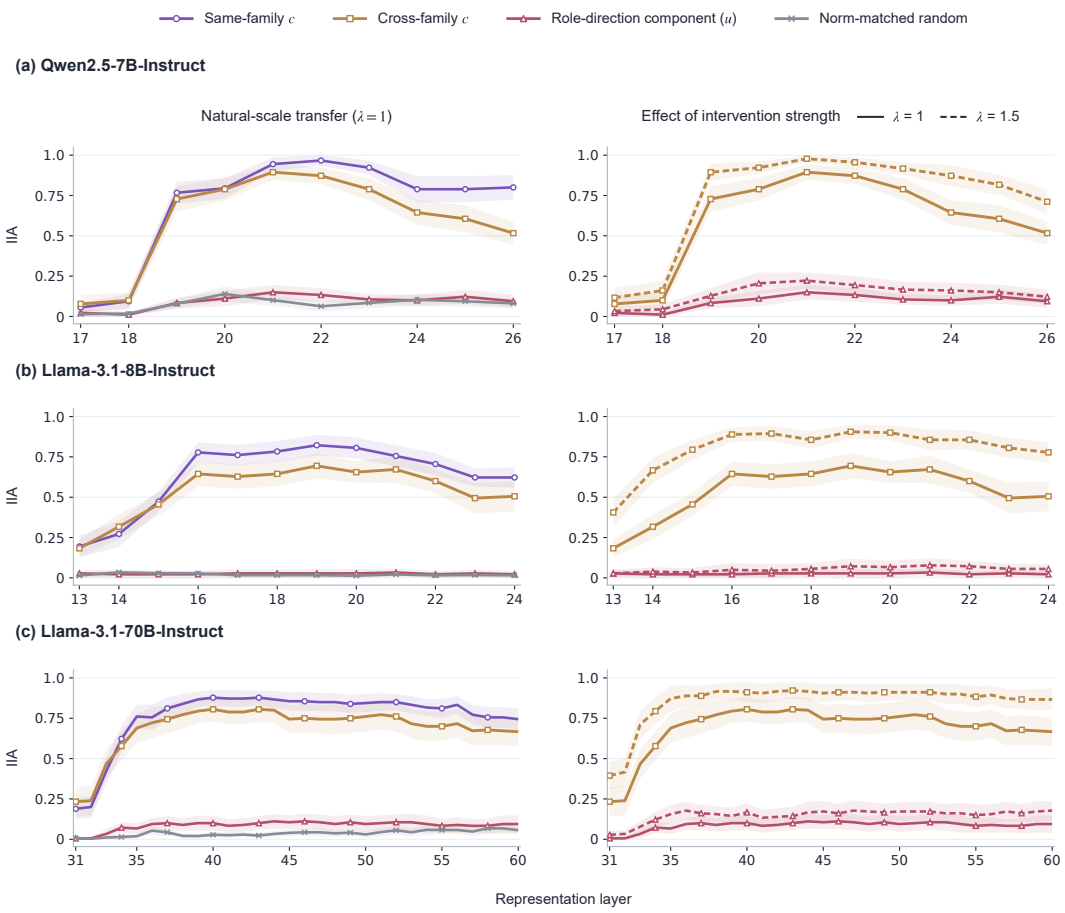

Figure 35: **Conflict Transfer across families.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. Components constructed under the `Invalid` rule are added to within-boundary bases under the `Unknown` rule. IIA measures production of the base-prescribed `Unknown`. Left: same-family c^ℓ, cross-family c^ℓ and u^ℓ, and the norm-matched random control at $\lambda = 1$. Right: cross-family c^ℓ and u^ℓ at $\lambda = 1$ (solid) and $\lambda = 1.5$ (dashed). The horizontal axis gives the selected Response intervention layers. Shaded bands indicate approximate 95% confidence intervals.

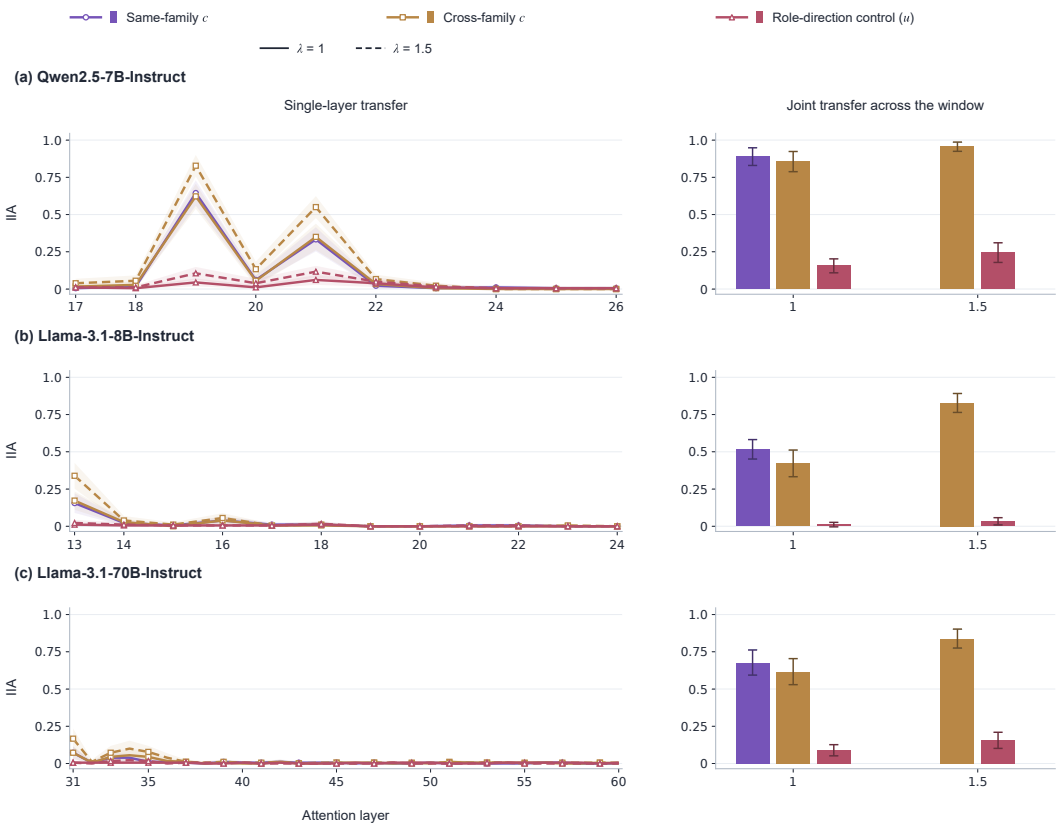

Figure 36: **Query-Induced Attention Writes at Response.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. IIA measures production of the base-prescribed Unknown after adding the attention-output components. Left: independent single-layer additions, with solid and dashed curves denoting $\lambda = 1$ and $\lambda = 1.5$, respectively. Right: joint additions of the layer-specific components across the selected interval in a single run. Same-family c^ℓ provides the reference at $\lambda = 1$; cross-family c^ℓ and the role-direction control u^ℓ are tested at both scales. The horizontal axis on the left gives the attention layer. Shaded bands and error bars indicate approximate 95% confidence intervals.

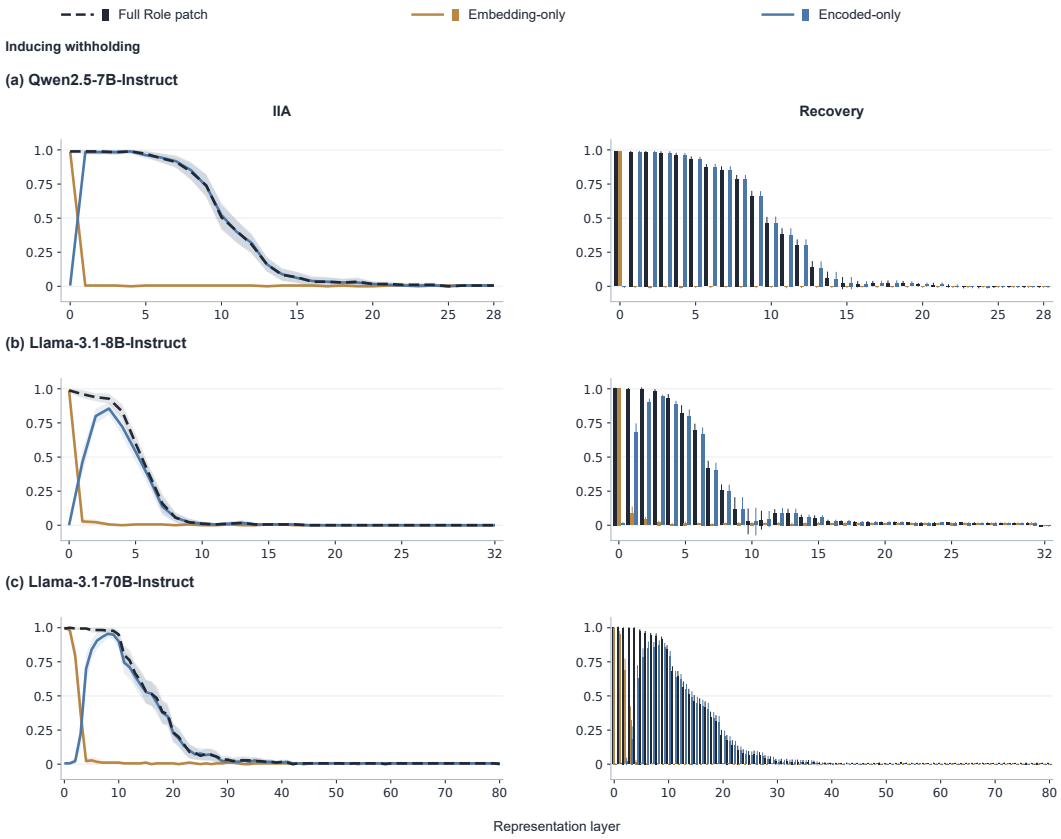

Figure 37: **Role Representation Decomposition toward withholding.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is beyond-boundary and the base is within-boundary, targeting `Unknown`. IIA curves (left) and normalized logit-margin recovery bars (right) compare full Role patching with Embedding-only and Encoded-only interventions. The horizontal axis gives the intervention layer for all three conditions. Shaded bands and error bars indicate approximate 95% confidence intervals.

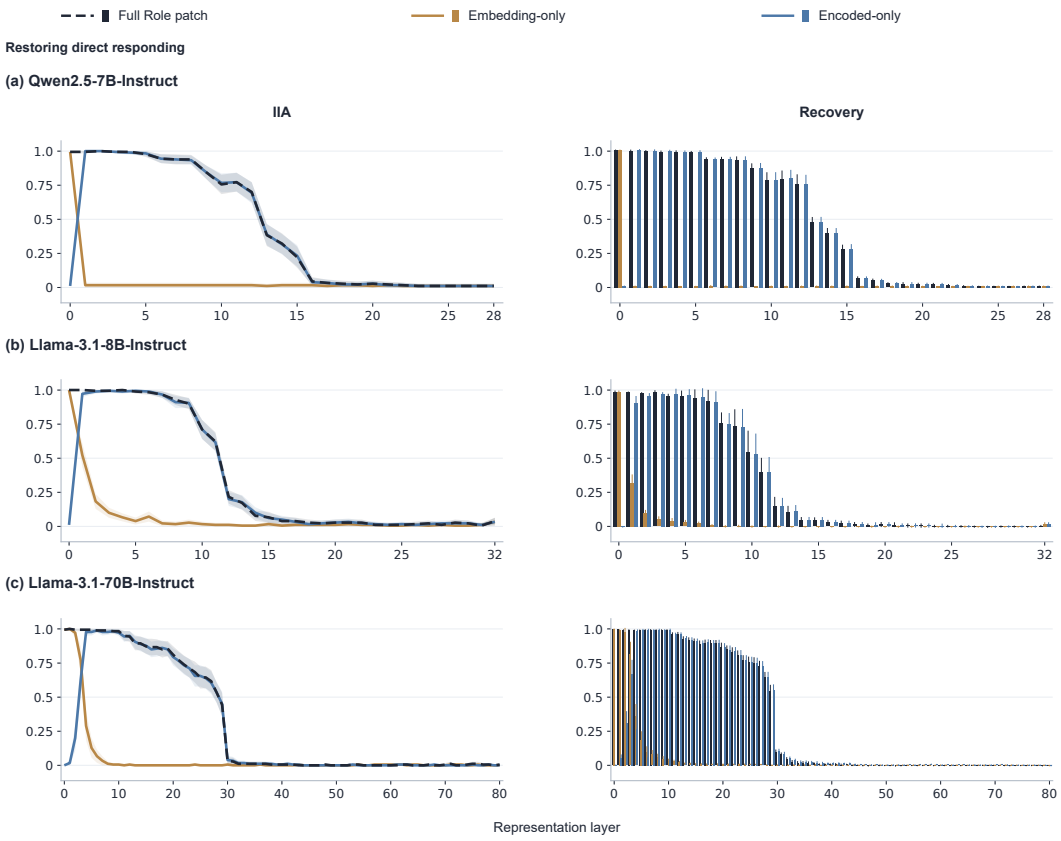

Figure 38: **Role Representation Decomposition toward direct responding.** Panels (a), (b), and (c) show Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, respectively. The source is within-boundary and the base is beyond-boundary, testing the restoration of direct responding. IIA curves (left) and normalized logit-margin recovery bars (right) compare full Role patching with Embedding-only and Encoded-only interventions. The horizontal axis gives the intervention layer for all three conditions. Shaded bands and error bars indicate approximate 95% confidence intervals.

Cross-Mapping Interventions, and Conflict Transfer are presented in Figures 47–48, 49–52, and 53, respectively. Query-Induced Attention Writes and Role Representation Decomposition are shown in Figure 54 and Figures 55–56, respectively. These results reproduce the key causal signatures identified using the original role sets.

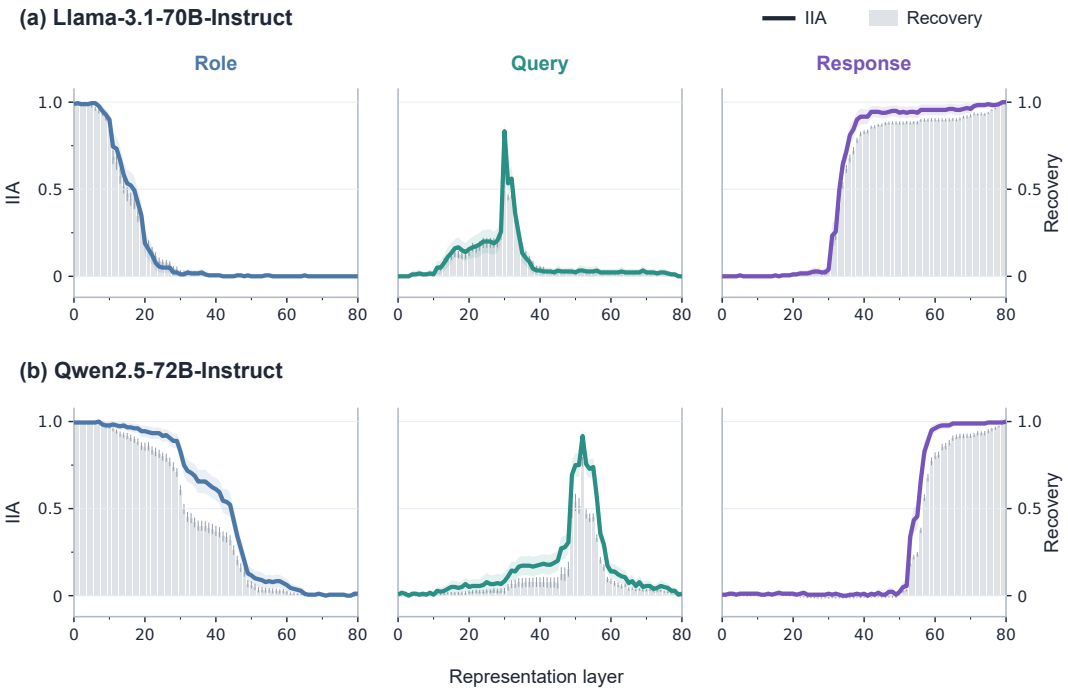

Figure 39: **Causal effect localization on the confirmation set.** The source is beyond-boundary and the base is within-boundary. Columns show Role, Query, and Response interventions. Solid curves show IIA for producing `Unknown` (left axes), and bars show normalized logit-margin recovery (right axes). The horizontal axis gives the intervention layer. Shaded bands and error bars indicate approximate 95% confidence intervals.

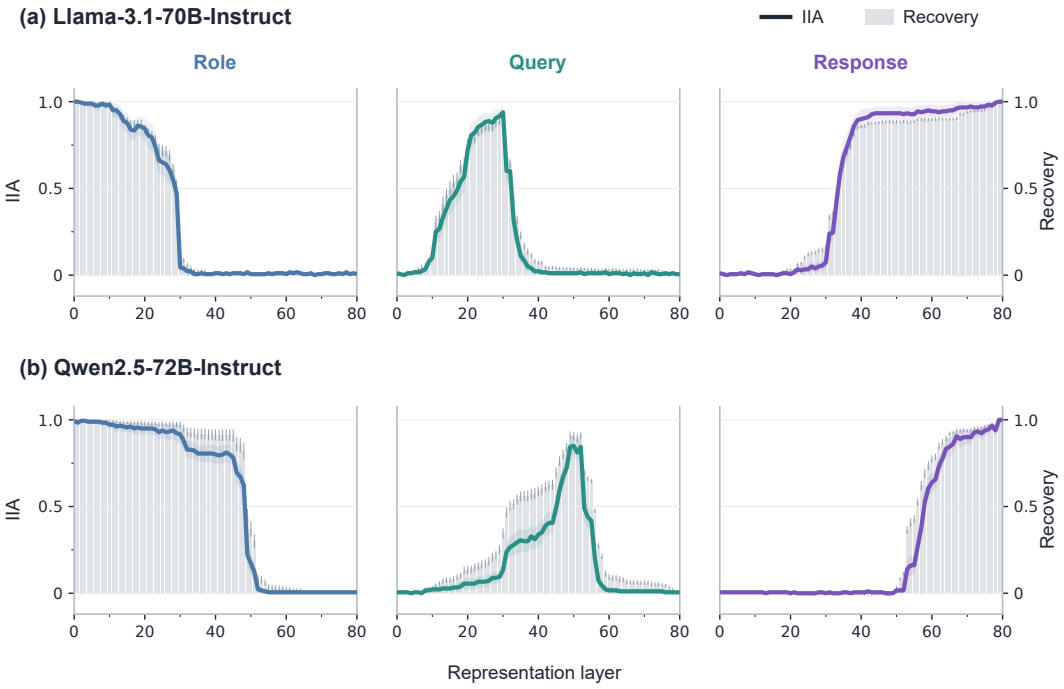

Figure 40: **Causal effect localization on the confirmation set.** The source is within-boundary and the base is beyond-boundary. Columns show Role, Query, and Response interventions. Solid curves show IIA for restoring direct responding (left axes), and bars show normalized logit-margin recovery (right axes). The horizontal axis gives the intervention layer. Shaded bands and error bars indicate approximate 95% confidence intervals.

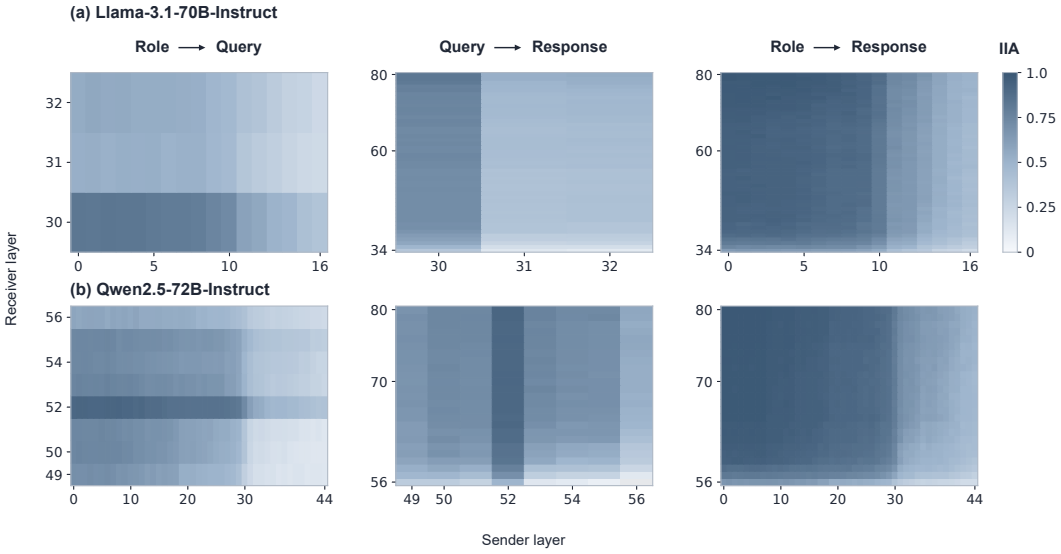

Figure 41: **Cross-Site Representation Replay on the confirmation set toward withholding: IIA.** The source is beyond-boundary and the base is within-boundary. Heatmaps show IIA for producing Unknown. Columns show Role-to-Query, Query-to-Response, and Role-to-Response replay; horizontal and vertical axes give the sender and receiver layers, respectively. Gray cells are not tested.

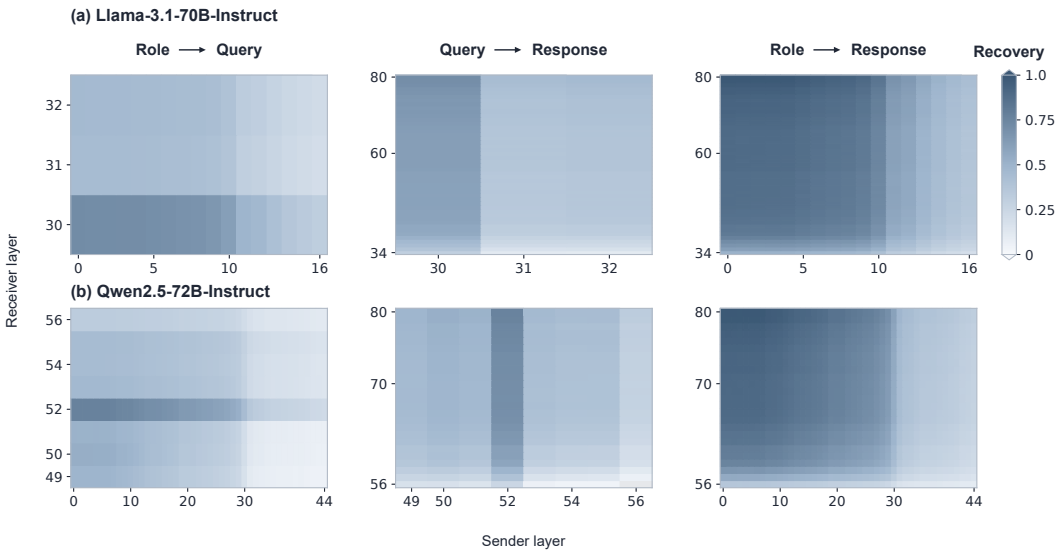

Figure 42: **Cross-Site Representation Replay on the confirmation set toward withholding: recovery.** The source is beyond-boundary and the base is within-boundary. Heatmaps show normalized logit-margin recovery. Columns show Role-to-Query, Query-to-Response, and Role-to-Response replay; horizontal and vertical axes give the sender and receiver layers, respectively. Gray cells are not tested; the upper extension of the color bar denotes recovery above 1.

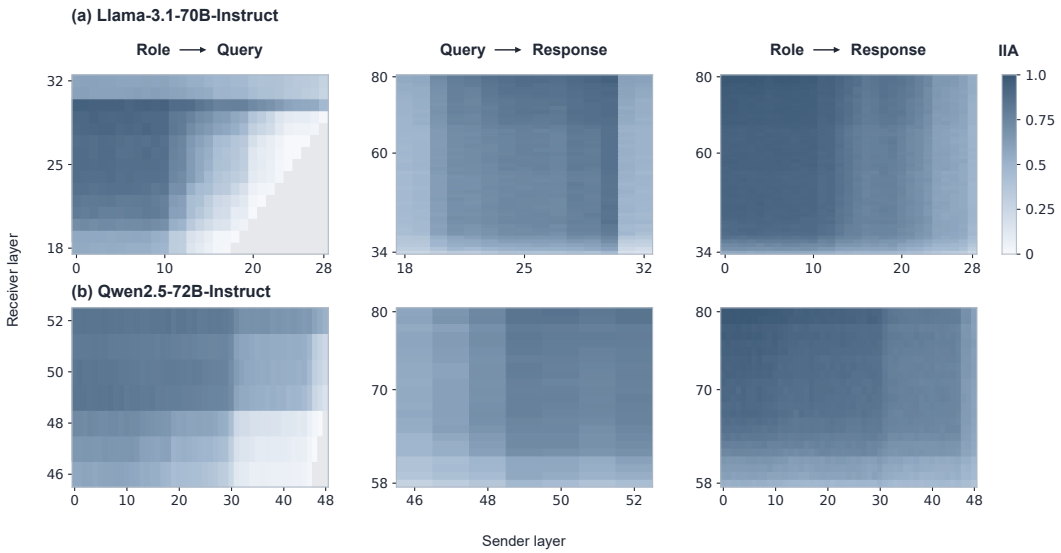

Figure 43: **Cross-Site Representation Replay on the confirmation set toward direct responding: IIA.** The source is within-boundary and the base is beyond-boundary. Heatmaps show IIA for restoring direct responding. Columns show Role-to-Query, Query-to-Response, and Role-to-Response replay; horizontal and vertical axes give the sender and receiver layers, respectively. Gray cells are not tested.

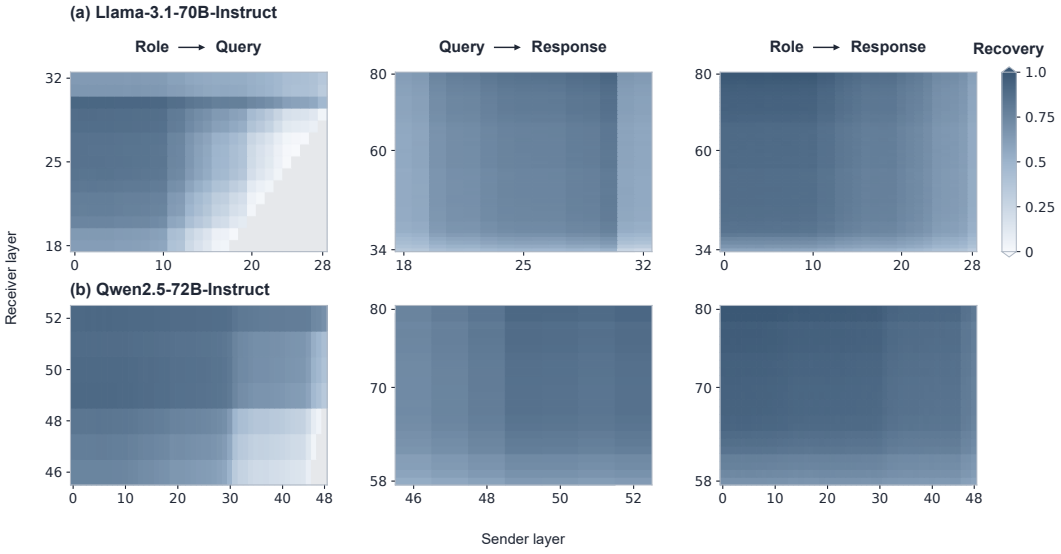

Figure 44: **Cross-Site Representation Replay on the confirmation set toward direct responding: recovery.** The source is within-boundary and the base is beyond-boundary. Heatmaps show normalized logit-margin recovery. Columns show Role-to-Query, Query-to-Response, and Role-to-Response replay; horizontal and vertical axes give the sender and receiver layers, respectively. Gray cells are not tested; the upper extension of the color bar denotes recovery above 1.

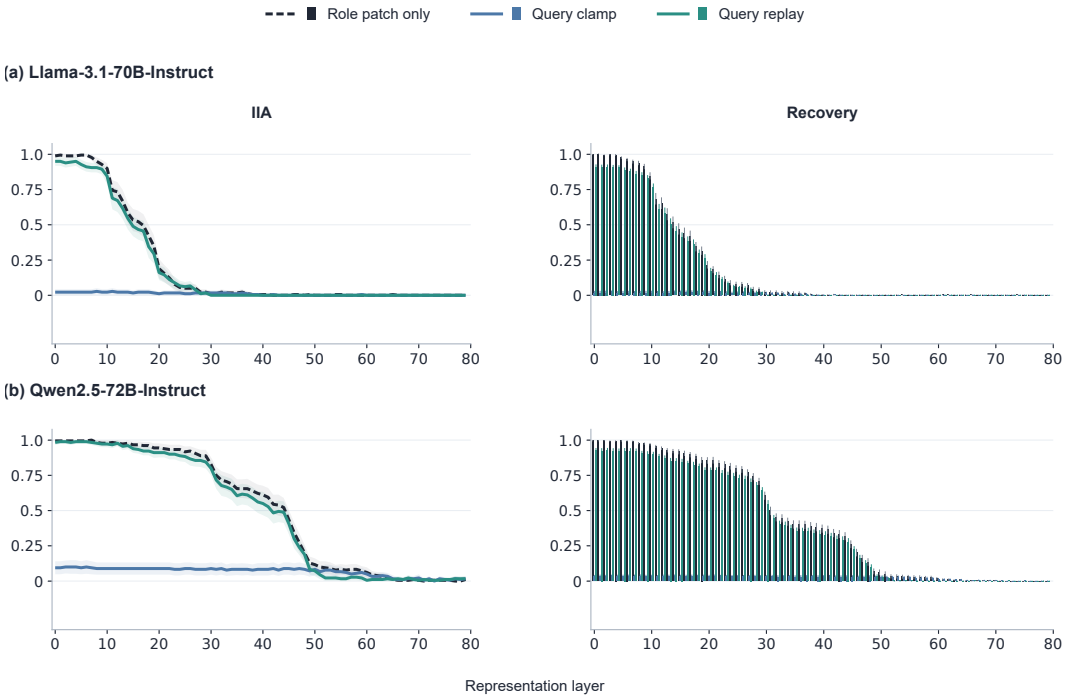

Figure 45: **Query mediation on the confirmation set.** Role patching uses a beyond-boundary source and a within-boundary base, targeting `Unknown`. IIA curves (left) and normalized logit-margin recovery bars (right) compare Role patching alone, Query clamp, and Query replay. Clamp results report the remaining effect. The horizontal axis gives the Role-patch layer. Shaded bands and error bars indicate approximate 95% confidence intervals.

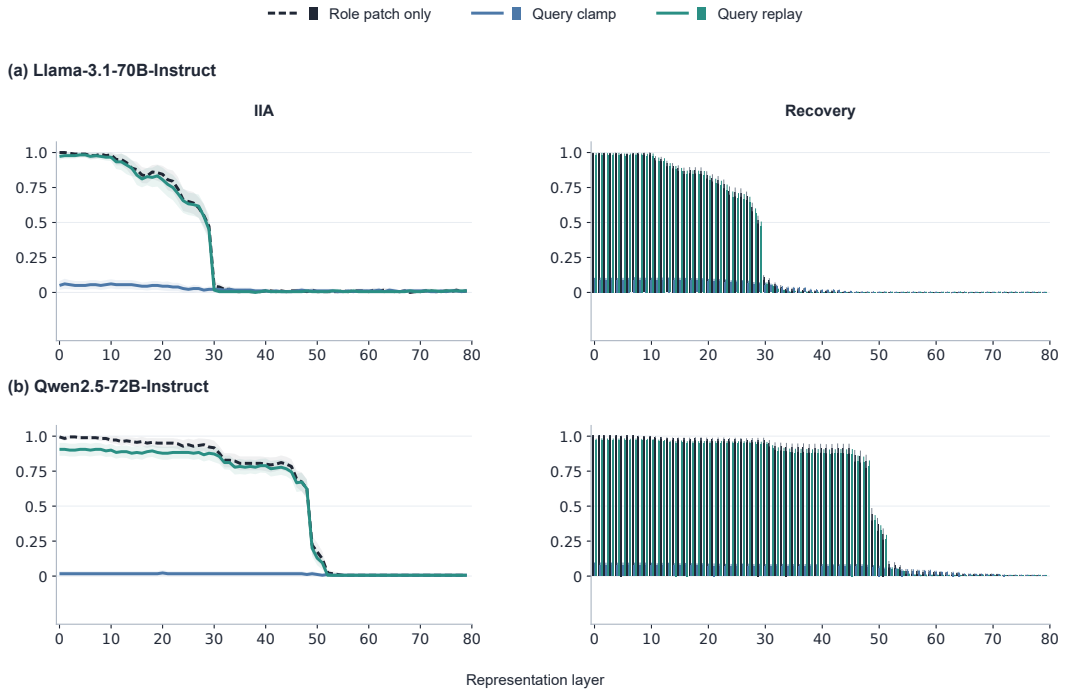

Figure 46: **Query mediation on the confirmation set.** Role patching uses a within-boundary source and a beyond-boundary base, targeting direct responding. IIA curves (left) and normalized logit-margin recovery bars (right) compare Role patching alone, Query clamp, and Query replay. Clamp results report the remaining effect. The horizontal axis gives the Role-patch layer. Shaded bands and error bars indicate approximate 95% confidence intervals.

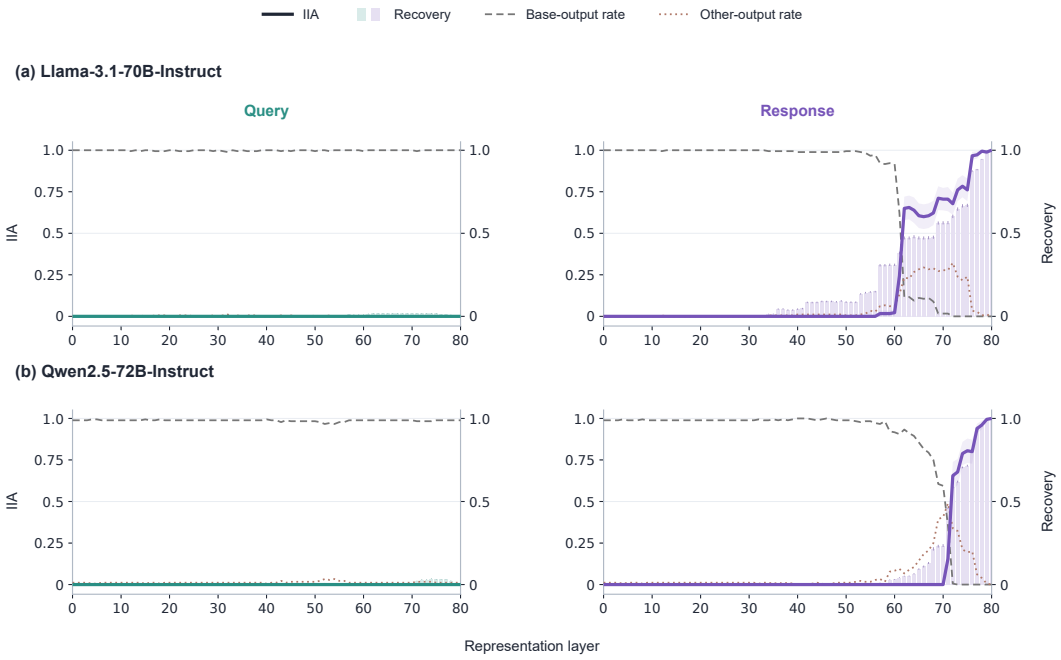

Figure 47: **Response mapping on the confirmation set: `Unknown` target.** The source and base share a beyond-boundary role–query instance, prescribing `Unknown` and `Invalid`, respectively. For Query and Response interventions, solid curves show IIA for `Unknown`; dashed and dotted curves give the base-output and other-output rates, respectively (left axes). Bars show normalized recovery of the `Unknown`–`Invalid` logit margin (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

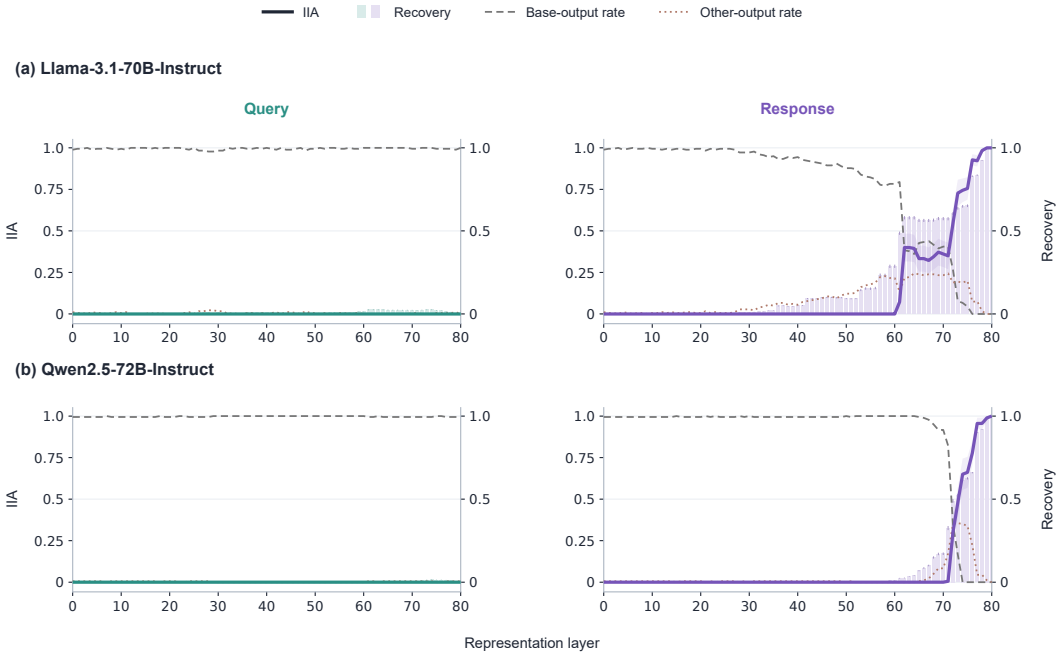

Figure 48: **Response mapping on the confirmation set: `Invalid` target.** The source and base share a beyond-boundary role–query instance, prescribing `Invalid` and `Unknown`, respectively. For Query and Response interventions, solid curves show IIA for `Invalid`; dashed and dotted curves give the base-output and other-output rates, respectively (left axes). Bars show normalized recovery of the `Invalid`–`Unknown` logit margin (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.



Figure 49: **Cross-Mapping Interventions on the confirmation set.** The beyond-boundary source prescribes `Invalid` for withholding, and the within-boundary base prescribes `Unknown`. Solid curves show IIA for either withholding word; dashed and dash-dotted curves show source- and base-prescribed response rates, respectively (left axes). Bars show normalized recovery of the logit margin between the original source withholding and base direct outputs (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

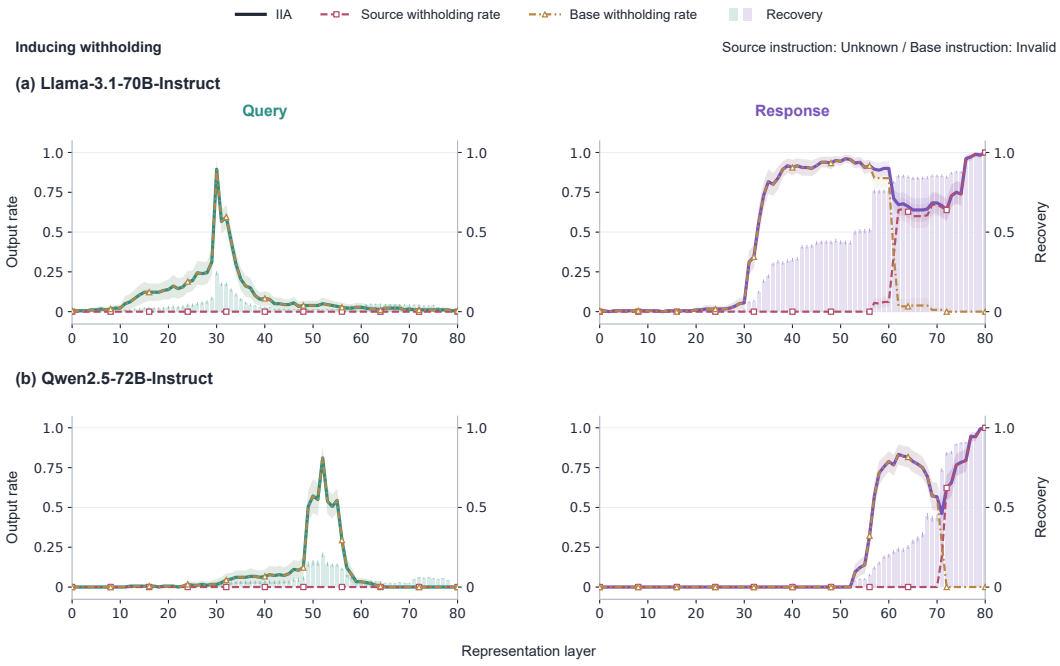

Figure 50: **Cross-Mapping Interventions on the confirmation set.** The beyond-boundary source prescribes `Unknown` for withholding, and the within-boundary base prescribes `Invalid`. Solid curves show IIA for either withholding word; dashed and dash-dotted curves show source- and base-prescribed response rates, respectively (left axes). Bars show normalized recovery of the logit margin between the original source withholding and base direct outputs (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

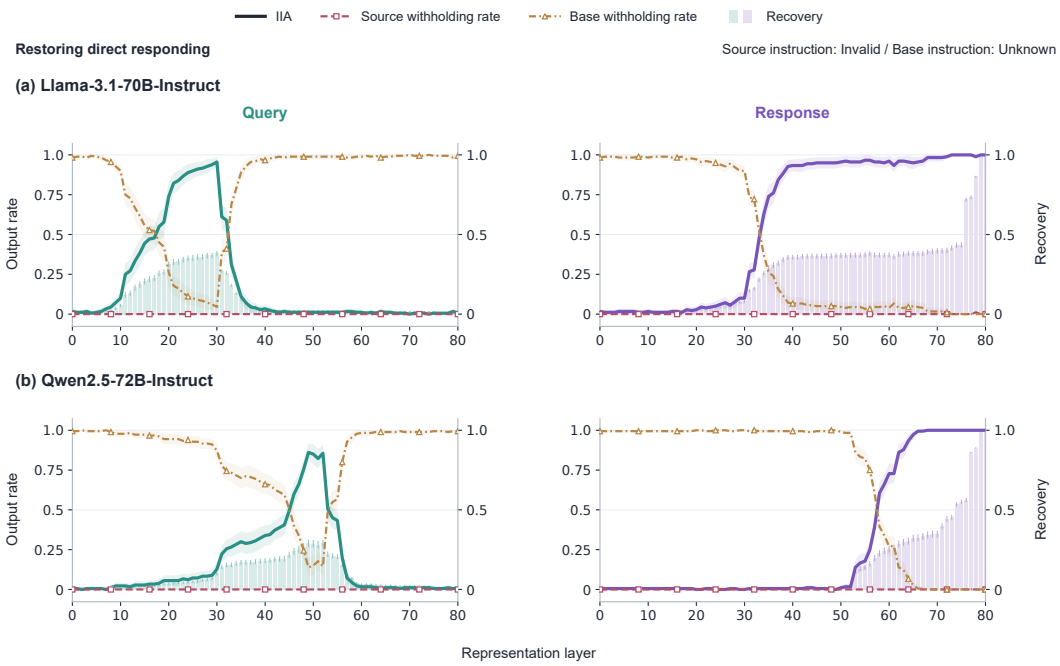

Figure 51: **Cross-Mapping Interventions on the confirmation set.** The within-boundary source prescribes `Invalid` for withholding, and the beyond-boundary base prescribes `Unknown`. Solid curves show IIA for outputs other than the two withholding words; dashed and dash-dotted curves show source- and base-prescribed withholding rates, respectively (left axes). Bars show normalized recovery of the logit margin between the original source direct and base withholding outputs (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

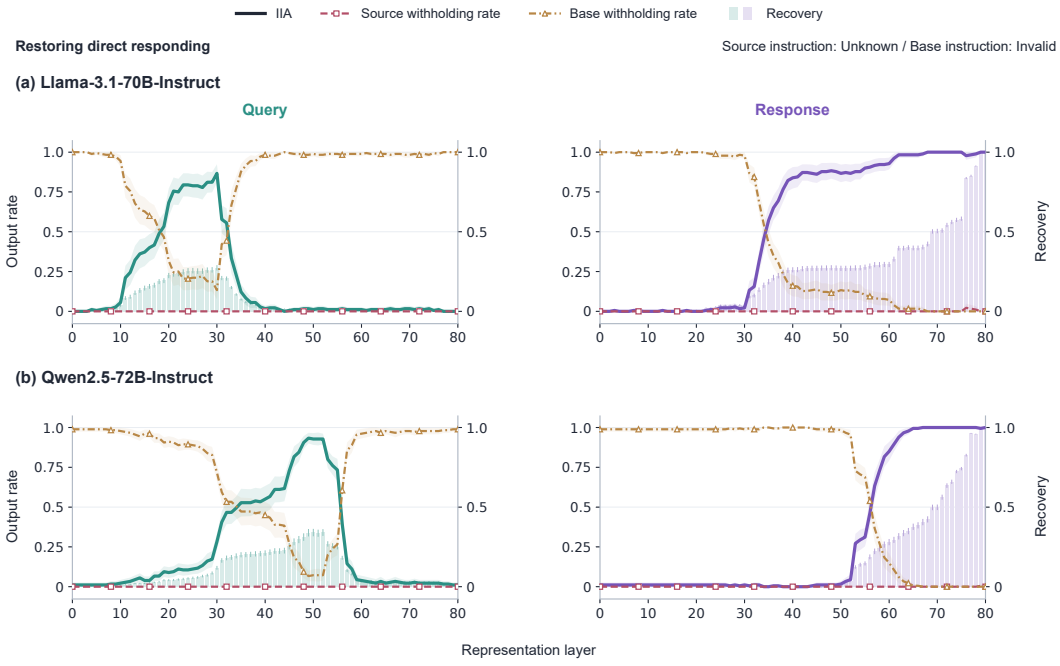

Figure 52: **Cross-Mapping Interventions on the confirmation set.** The within-boundary source prescribes `Unknown` for withholding, and the beyond-boundary base prescribes `Invalid`. Solid curves show IIA for outputs other than the two withholding words; dashed and dash-dotted curves show source- and base-prescribed withholding rates, respectively (left axes). Bars show normalized recovery of the logit margin between the original source direct and base withholding outputs (right axes). Shaded bands and error bars indicate approximate 95% confidence intervals.

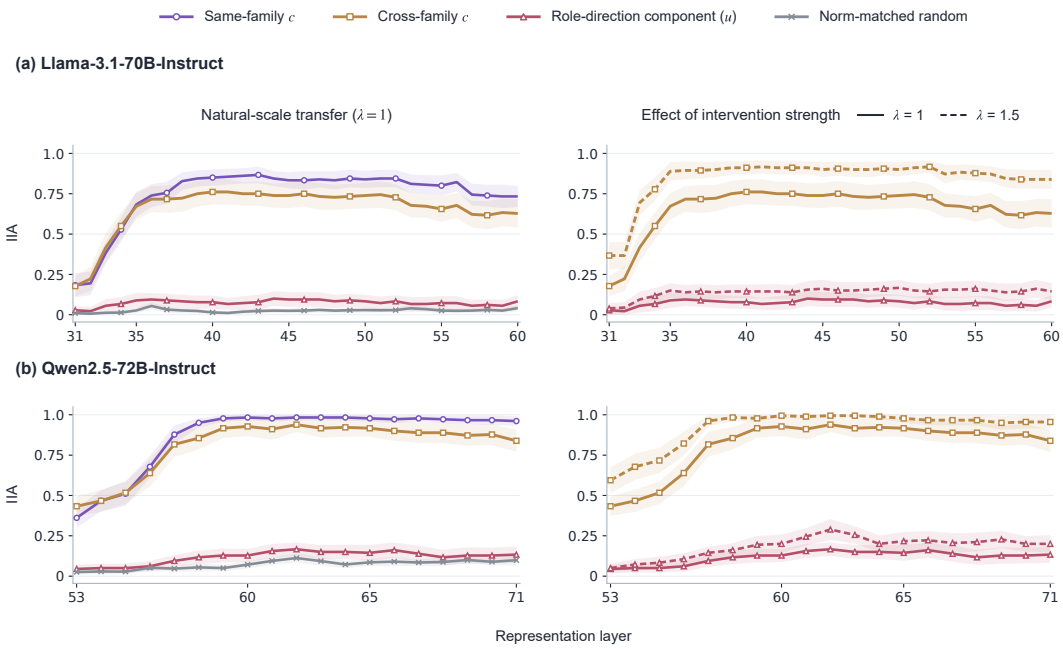

Figure 53: **Conflict Transfer on the confirmation set.** Components constructed under the `Invalid` rule are added to within-boundary bases under the `Unknown` rule; IIA measures production of `Unknown`. Left: same-family c^ℓ, cross-family c^ℓ and u^ℓ, and the norm-matched random control at $\lambda = 1$. Right: cross-family c^ℓ and u^ℓ at $\lambda = 1$ (solid) and $\lambda = 1.5$ (dashed). The horizontal axis gives the selected Response layers; shaded bands indicate approximate 95% confidence intervals.

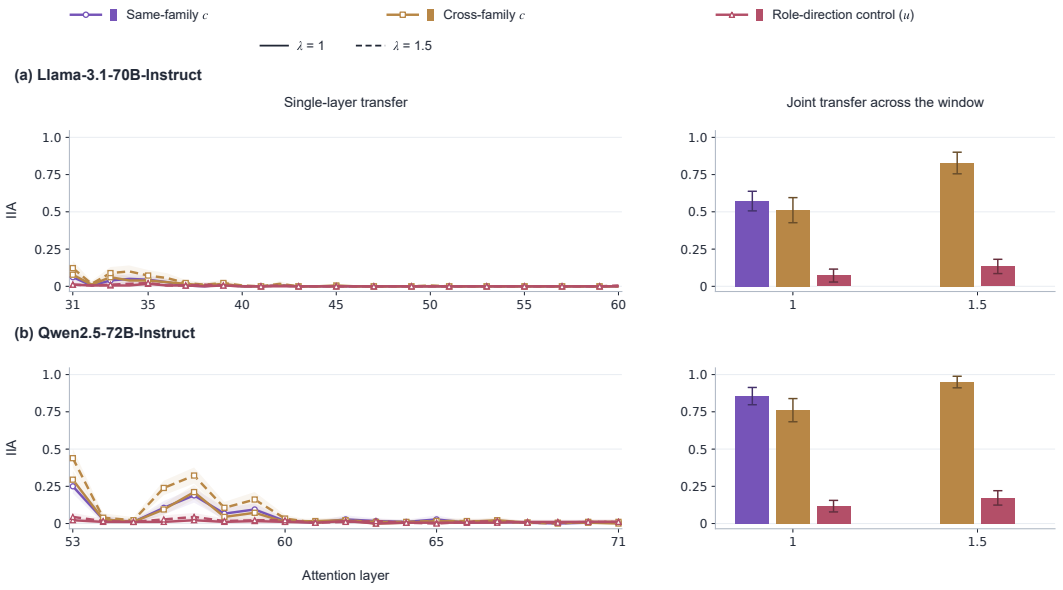

Figure 54: **Query-Induced Attention Writes on the confirmation set.** IIA measures production of the base-prescribed `Unknown`. Left: independent single-layer additions, with solid and dashed curves denoting $\lambda = 1$ and $\lambda = 1.5$, respectively. Right: joint additions of the layer-specific components across the selected interval in one run. Same-family c^ℓ is the reference at $\lambda = 1$; cross-family c^ℓ and u^ℓ are tested at both scales. Shaded bands and error bars indicate approximate 95% confidence intervals.

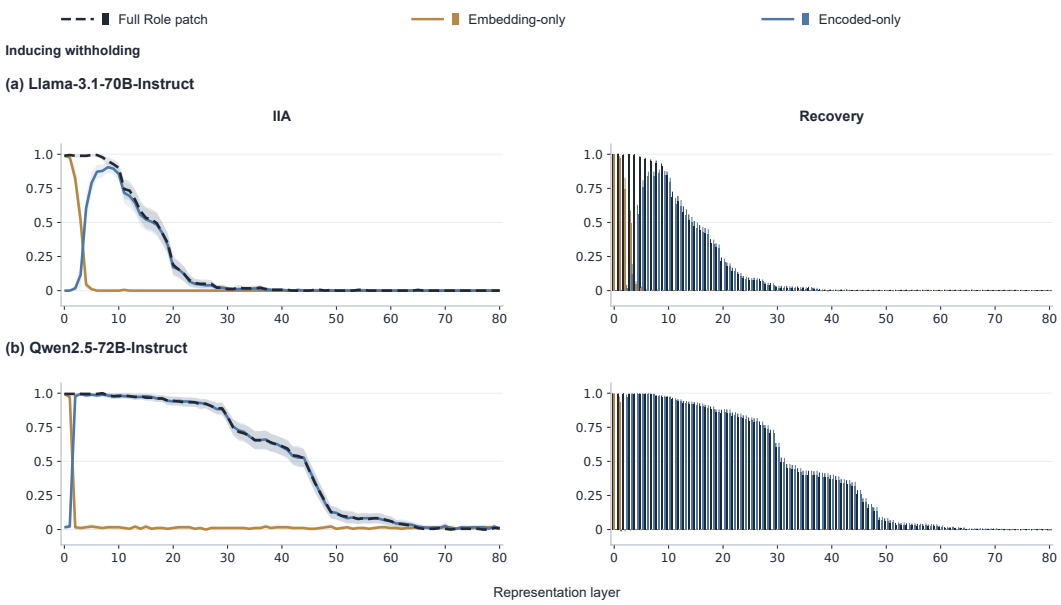

Figure 55: **Role Representation Decomposition on the confirmation set toward withholding.** The source is beyond-boundary and the base is within-boundary. IIA measures production of Unknown. IIA curves (left) and normalized logit-margin recovery bars (right) compare full Role patching, Embedding-only, and Encoded-only interventions. The horizontal axis gives the intervention layer for all conditions. Shaded bands and error bars indicate approximate 95% confidence intervals.

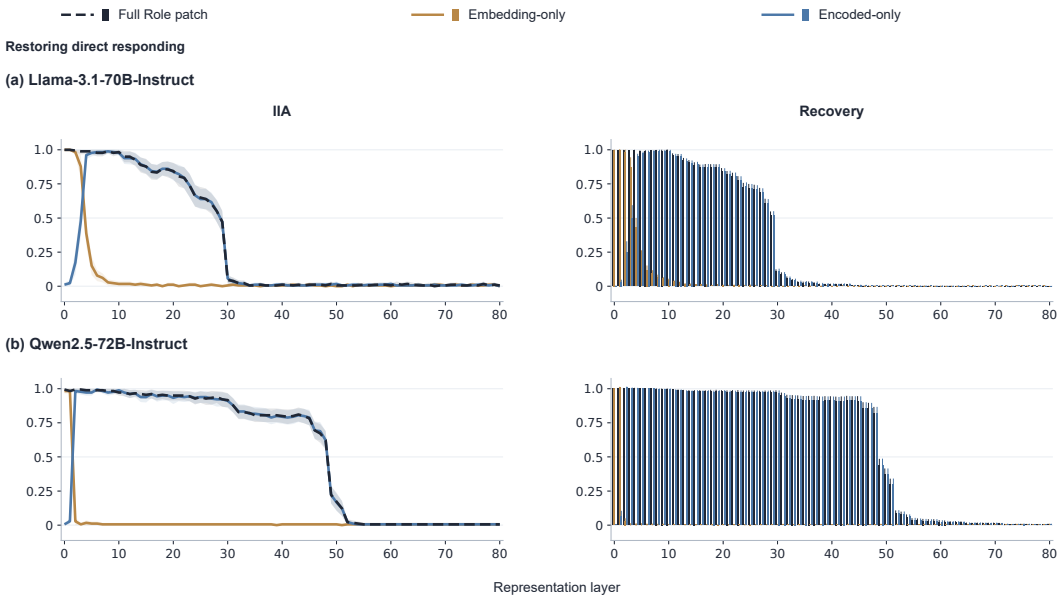

Figure 56: **Role Representation Decomposition on the confirmation set toward direct responding.** The source is within-boundary and the base is beyond-boundary. IIA measures restoration of direct responding. IIA curves (left) and normalized logit-margin recovery bars (right) compare full Role patching, Embedding-only, and Encoded-only interventions. The horizontal axis gives the intervention layer for all conditions. Shaded bands and error bars indicate approximate 95% confidence intervals.