A MECHANISTIC ANALYSIS OF IN-CONTEXT LEARNING AND OUT-OF-CONTEXT HALLUCINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can infer and follow specific rules through in-context learning (ICL), yet may fail to remain faithful to provided evidence, giving rise to out-of-context (OOC) hallucination. Existing work has largely investigated the underlying mechanisms of ICL and OOC hallucination along separate lines. However, this separation obscures the mechanistic basis for these seemingly opposing outcomes. In this work, we study this contrast through the unified view of functional roles in dynamic context. The key observation is that role information can be internally represented without reliably constraining output. We derive a first-order decomposition to characterize the role effect and identify two necessary conditions for role span contributions to become consequential for generation. Causal analysis further shows that role effects act through key token choices. These findings support a common mechanistic view of ICL and OOC hallucination: each functional role defines the pattern of output decisions over which role span contributions must remain effective, and differences in this organization account for the seemingly opposing outcomes.

1 INTRODUCTION

A central strength of LLMs is the ability to condition generation on dynamic instructions. This capability is exemplified by in-context learning (ICL), where models can infer and follow specific rules supplied in the prompt (Brown et al., 2020; Dong et al., 2024). The same dependence on context, however, also exposes a major reliability challenge. Models may fail to remain faithful to the provided evidence and even produce unsupported claims, a failure mode we refer to as out-of-context (OOC) hallucination (Huang et al., 2025b; Wu et al., 2024). This contrast raises a natural question: why can the same model reliably follow a prompt-specified rule, yet fail to remain faithful to evidence that is likewise supplied in context? Although existing work studies the mechanisms underlying ICL and OOC hallucination separately (Hendel et al., 2023; Todd et al., 2024; Yu et al., 2024), the mechanistic origin of this contrast remains unclear.

Fundamentally, dynamic context serves different functional roles in generation: rules specify how to respond, whereas evidence specifies what factual content the response should be grounded in. The same content can be interpreted differently according to its function in the current prompt. The same sentence may act as a rule, a distractor, or ordinary background across tasks (Shi et al., 2023). While contemporary LLMs receive no explicit supervision for these roles, recent mechanistic studies of ICL suggest that task structure can nevertheless be inferred from context (Cho et al., 2025).

In this paper, we present a counter-intuitive finding: the functional roles can be internally represented, yet fail to reliably constrain the resulting response. We decompose this distinction into two questions: (1) how functional roles are internally represented, and (2) how role information affects LLM generation. Despite their different outcomes, ICL and OOC hallucination can be analyzed through a unified framework focusing on the role effect, while preserving the distinct functions of rules and evidence. Accordingly, the role effect depends on how role information influences generation as decoding proceeds.

The effect of dynamic context is inherently non-uniform (An et al., 2024). The prefix can establish the direction of the continuation and persistently influence subsequent predictions. For instance, rules can shape how the response is organized, whereas evidence can determine which factual content is carried forward (Xie et al., 2024; Du et al., 2024; Wu et al., 2025). We term the corresponding

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.

role span's effect on the relevant next-token predictions the role span contribution. Our experiments show that a role can guide the continuation only when its role span contribution continues to affect next-token predictions at the relevant decoding step. From this perspective, we show that ICL and OOC hallucination share the same contextual-influence mechanism, while it takes effect at different types of decisions in the two settings. Interested readers can refer to Appendix A for a thorough discussion of related work.

Statement of Contribution. Our contributions are threefold. First, we establish a unified account of contextual influence in ICL and OOC hallucination by distinguishing internal role representation from its effect on generation. Second, we derive a first-order decomposition and identify two necessary conditions for role span contributions to exert a causal influence on generation. Third, we show that the distinct outcomes of the two settings can be explained by different decision-level realizations of a shared role span contribution mechanism.

2 PRELIMINARIES

Problem formulation. We study ICL and OOC hallucination as two autoregressive generation regimes where the context plays different functional roles. In ICL, the context supplies a rule or instruction for answering the query. In OOC hallucination, the context supplies evidence against which the generated response is evaluated.

Functional roles. We model both ICL and OOC hallucination under the same autoregressive decoding interface. Let C denote the context, x the query, and $y = (y_1, \ldots, y_T)$ a realized response. At decoding step t, the model defines a next-token distribution $p_\theta(\cdot \mid C, x, y_{<t})$. We index the two functional roles, rule and evidence, by $f \in \{r, e\}$, and use C_f to denote the corresponding context span.

Role span positions. Since a role f is instantiated by role-bearing positions in the prompt, we focus on the corresponding set of input positions. Given an input P, we define $\mathcal{S}_f(P)$ as the set of input-side token positions instantiating role f. Specifically, $\mathcal{S}_r(P)$ denotes the rule span, and $\mathcal{S}_e(P)$ denotes the evidence span. These positions provide the input-side span used in the matched comparisons below.

Matched residual response. We use matched role comparisons to isolate the effect of role f. For each role context C_f, we construct a matched counterfactual context $\widetilde{C}_f$. The counterfactual context changes only the role-bearing component while holding the remaining prompt fixed. Based on this, we analyze the two task forms through the same matched comparison, where the changed component corresponds to the rule span in ICL and the evidence span in OOC hallucination. Given a sample y, we define matched processed inputs $P^f(y)$ and $\widetilde{P}^f(y)$, constructed from $C_f \oplus x \oplus y$ and $\widetilde{C}_f \oplus x \oplus y$, where $\oplus$ denotes sequence concatenation. Construction details are provided in Appendix C.1. Motivated by recent work that studies LLM behavior through hidden states (Hendel et al., 2023; Todd et al., 2024), we carry this comparison into the residual stream. Specifically, we consider token positions i, including those in the role span $\mathcal{S}_f(P^f(y))$. Let $h_\ell(P, i) \in \mathbb{R}^d$ denote the residual-stream representation at layer ℓ and position i when the model processes input P. The induced representation difference is $\Delta h^f_\ell(i; y) = h_\ell(P^f(y), i) - h_\ell(\widetilde{P}^f(y), i)$. With this sign convention, $\Delta h^f_\ell(i; y)$ points from the matched counterfactual state to the original role instantiation. This signed difference provides the residual response object used in the representation analysis below.

3 ROLE SPANS INDUCE FUNCTIONAL RESIDUAL RESPONSES

ICL and OOC hallucination differ in their empirical outcomes, yet both rely on dynamic context within the same autoregressive decoding framework. Despite the shared structure, context serves distinct functional roles in the two settings. This raises a natural question: ***how does the functional role organize the model's internal state***? We address this through matched role comparisons in the residual stream, linking the role distinction to its internal representational form. We first characterize the organization of the induced state changes and then examine their relation to the decoding process.

3.1 Structured Residual Responses from Role Spans

The same sentence can serve different functional roles, such as an instruction, evidence, or distractor. To examine how this distinction appears in internal states, we keep the span content fixed while changing only its function. Specifically, we construct multiple functional framings for each span and train a linear readout to distinguish the corresponding roles, following Belinkov (2022). We find that the readout generalizes consistently across prompt formulations, indicating a stable functional distinction in internal states.

We next examine how functional role structures internal states. To test this, for each role f, we fit the top 10 PCA directions on one half of the samples and measure the fraction of Δh_ℓ^f energy captured on the other half. Ideally, a high captured fraction would indicate shared organization, whereas isotropic differences would remain near the random baseline. Figure 1 shows that the energy concentration remains well above the baseline. Rule responses exhibit strong concentration, whereas evidence responses are more distributed. These results show that the residual responses induced by functional role exhibit systematic low-dimensional organization.

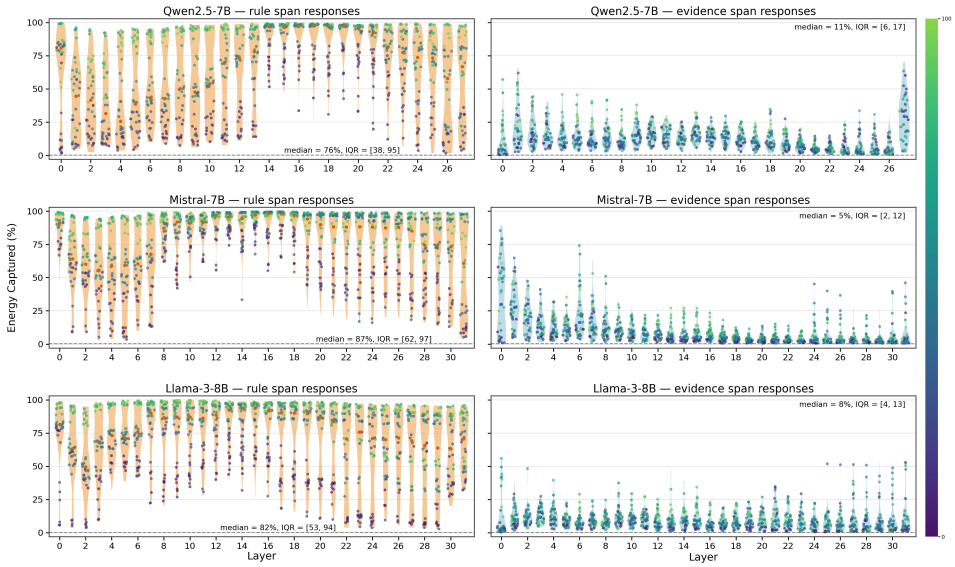


Figure 1: **Structured residual responses**. Left panels show rule span responses, and right panels show evidence span responses. Each violin shows the fraction of Δh energy captured by a 10-dimensional PCA subspace; points are colored by the percentile of $\|\Delta h\|$. The dashed gray line denotes the expected random subspace baseline $k/H \approx 0.2\%$, with $k = 10$ PCA directions and residual dimension.

3.2 Residual Responses Toward Role-Consistent Generation

The low-dimensional organization observed above shows that role spans can induce structured representational change during inference. Since the residual stream connects prompt processing to token prediction, we next examine how these changes relate to the output process. To make this relation concrete, we consider two empirical implications: (1) **The dominant directions of the residual subspace should align with the unembedding directions of outputs.** This follows from the fact that the LLM maps the residual representation to token logits through the final normalization and unembedding matrix W_U. Here, the structured Δh should appear as a larger projection onto the output directions. (2) **The representation change induced by a specific role span should preserve relevant role information.** Since the residual stream connects intermediate representations with downstream computation, the role information represented by Δh should be maintained in a form that later blocks can read, transform, and propagate.

To examine the first empirical implication, we project Δh through W_U, measuring its projection and logit shift toward the output. We use a matched non-role perturbation as the comparison group. Figure 2 shows that, across all three models, role responses are more strongly aligned with the corresponding output directions, which results in larger logit shifts. Specifically, rule responses

show clearer positive projection, while evidence responses separate more through the logit shift. By contrast, the comparison group remains weaker in both projection and logit shift. These results show that, during the decoding process, structured Δh appears as a directional logit shift toward the output direction indicated by the role span.

To examine the second empirical implication, we evaluate behavior recovery in states induced by the matched counterfactual context; details are provided in Appendix C.2. As shown in Table 1, matched role responses restore the original behavior associated with the role, showing that the injected residual response remains usable by subsequent computation. Moreover, recovery remains low for the comparison groups since they lack the role conditioned state required for recovery. These results indicate that the induced representational change carries transferable role information that remains usable by downstream computation.

Our findings suggest that functional roles in ICL and OOC hallucination can be characterized in the representation space. Rule and evidence spans induce a common form of structured residual response that carries role-specific information forward for downstream computation.

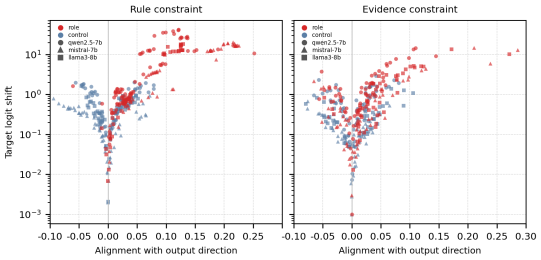


Figure 2: **Final layer decoding projection of Δh.** Blue points denote the comparison group.

Table 1: **Behavior recovery rates**.

Injected response	Qwen	Mistral	Llama	Avg.
Rule				
matched role response	**0.94**	**1.00**	**1.00**	**0.98**
no role content	0.17	0.08	0.00	0.08
mismatched role response	0.10	0.31	0.05	0.15
Evidence				
matched role response	**0.71**	**0.80**	**0.93**	**0.81**
no role content	0.44	0.20	0.37	0.34
mismatched role response	0.35	0.25	0.04	0.21

4 Role Span Contributions in Generation

So far, we have characterized how role spans are represented in the residual stream, yet how a role span shapes LLM generation remains unclear. We therefore first formulate the effect of a role span on generation, then identify how role span information enters the LLM forward pass, and finally derive the necessary conditions for the role span to shift LLM generation.

4.1 Response Trajectory Score Shift

To formulate the effect of a role span, we fix a candidate response trajectory y and compare the trajectory under C_f and its matched counterfactual $\widetilde{C}_f$. At response position t, let $P_t^f(y)$ denote the input constructed from $C_f \oplus x \oplus y_{<t}$, and let $\widetilde{P}_t^f(y)$ denote the corresponding input under $\widetilde{C}_f$. The two matched inputs share the same query x and response prefix $y_{<t}$, differing only in the matched replacement of the role context. We write the stepwise score as $\psi_t(P) = \log p_\theta(y_t \mid P)$ for the fixed trajectory y. At each step, we compare the autoregressive scores under C_f and $\widetilde{C}_f$, which yields the matched score shift $\delta_t^f(y) = \psi_t(P_t^f(y)) - \psi_t(\widetilde{P}_t^f(y))$. Since the autoregressive probability of y is written as a product over response positions, the full log probability shift takes the form of accumulated stepwise differences across those positions. We formalize this quantity in the following definition.

Definition 1 (Response trajectory score shift)**.** For a fixed candidate response trajectory y, define the stepwise score shift as $\delta_t^f(y) = \psi_t(P_t^f(y)) - \psi_t(\widetilde{P}_t^f(y))$. The response trajectory score shift is $\Phi_f(y) = \sum_{t=1}^{T} \delta_t^f(y)$, which measures how the functional role span shapes the response trajectory under C_f and matched $\widetilde{C}_f$. Equivalently, using the autoregressive product form,

$$\begin{aligned}\Phi_f(y) &= \log p_\theta(y \mid C_f \oplus x) - \log p_\theta(y \mid \widetilde{C}_f \oplus x) \\ &= \sum_{t=1}^{T} \left[\log p_\theta(y_t \mid C_f \oplus x \oplus y_{<t}) - \log p_\theta(y_t \mid \widetilde{C}_f \oplus x \oplus y_{<t})\right].\end{aligned} \tag{1}$$

Here, $\Phi_f(y) > 0$ indicates that C_f yields a higher score than $\widetilde{C}_f$ does. In what follows, we relate the stepwise score shifts $\delta_t^f(y)$ to forward-pass quantities associated with the role span.

4.2 Role Span Contributions in the Forward Pass

We mainly consider decoder-only LLMs, where successive attention and MLP modules update the residual stream before the final unembedding. The role-induced responses observed above are therefore formed and propagated through these updates. To relate this process to the role span, we write the residual update explicitly. For a processed input P, the residual state at layer ℓ and position i is updated as $h_\ell(P,i) = h_{\ell-1}(P,i) + \mathrm{Attn}_\ell^{\mathrm{out}}(P,i) + \mathrm{MLP}_\ell^{\mathrm{out}}(P,i)$. This reveals that information carried by different roles is accumulated across depth. A difference induced by role f can therefore reflect both information inherited from previous layers and the current updates. Recall that the role span $\mathcal{S}_f(P)$ is defined by source token positions. Among the two update terms above, the attention output can be explicitly written as a sum over source token positions:

$$\mathrm{Attn}_\ell^{\mathrm{out}}(P,i) = \sum_a \sum_s \alpha_{i\leftarrow s}^{(a)}(P) V_\ell^{(a)}(P,s) W_O^{(a)} \tag{2}$$

where a indexes attention heads and s indexes source token positions in P that are visible to the target position i under the causal mask. Eq. (2) decomposes the attention update at i into contributions aggregated from source positions. The key observation is that the role span $\mathcal{S}_f(P)$ is also a set of input token positions. We can therefore isolate the contribution of the role span by retaining the terms with $s \in \mathcal{S}_f(P)$, giving

$$B_{\ell,i}^f(P) = \sum_a \sum_{s\in\mathcal{S}_f(P)} \alpha_{i\leftarrow s}^{(a)}(P) V_\ell^{(a)}(P,s) W_O^{(a)}. \tag{3}$$

We term $B_{\ell,i}^f(P)$ the role span contribution to the attention update at layer ℓ and position i. In decoder-only Transformers, attention directly transfers differences across token positions, whereas normalization and MLP blocks act only within individual positions. Appendix B.1 formalizes this distinction.

4.3 First-Order Decomposition of Score Shifts

The response trajectory score shift $\Phi_f(y)$ accumulates the stepwise shifts $\delta_t^f(y)$ across response positions. We therefore analyze each stepwise shift through selected role span contribution coordinates and quantify the first-order score change associated with their matched differences. To make this operational, we approximate the stepwise score by its first-order change with respect to the selected $B_{\ell,i}^f$ coordinates. Specifically, for each decoding step t, let $\mathcal{B}_{f,t} \subseteq \{(\ell,i)\}$ denote the role span contribution indices considered in the expansion. For $(\ell,i) \in \mathcal{B}_{f,t}$, define $\Delta B_{\ell,i,t}^f(y) = B_{\ell,i}^f(P_t^f(y)) - B_{\ell,i}^f(\widetilde{P}_t^f(y))$. Collecting these differences gives $\Delta B_{\mathcal{B},t}^f(y) = \left(\Delta B_{\ell,i,t}^f(y)\right)_{(\ell,i)\in\mathcal{B}_{f,t}}$. The vector $\Delta B_{\mathcal{B},t}^f(y)$, therefore, measures how the matched role comparison changes the selected role span contributions over $\mathcal{B}_{f,t}$.

Assumption 1 (Local score regularity)**.** *For a fixed response position t, let $\widetilde{b}$ and b^f denote the selected role span contribution vectors induced by $\widetilde{P}_t^f(y)$ and $P_t^f(y)$, respectively. Define $\Psi_t(b)$ as the score obtained by replacing, in layer order, the selected contributions under $\widetilde{P}_t^f(y)$ with the corresponding entries of b. By construction, $\Psi_t(\widetilde{b}) = \psi_t(\widetilde{P}_t^f(y))$; let $E_t^\perp = \psi_t(P_t^f(y)) - \Psi_t(b^f)$. We assume that Ψ_t is differentiable and has an L_t-Lipschitz gradient on the line segment from $\widetilde{b}$ to b^f, and write $\nabla_{B_{\ell,i}^f}\psi_t := \nabla_{b_{\ell,i}}\Psi_t(\widetilde{b})$.*

Under Assumption 1, a first-order expansion of Ψ_t from $\widetilde{b}$ to b^f yields the following decomposition.

Theorem 1 (First-order role span score decomposition)**.** *Suppose Assumption 1 holds. Then the score shift satisfies*

$$\delta_t^f(y) = \sum_{(\ell,i)\in\mathcal{B}_{f,t}} \left\langle \nabla_{B_{\ell,i}^f}\psi_t, \Delta B_{\ell,i,t}^f(y) \right\rangle + R_t, \tag{4}$$

where the gradients are evaluated at $\widetilde{b}$, and the remainder satisfies

$$|R_t| \le \frac{L_t}{2} \left\| \Delta B^f_{\mathcal{B},t}(y) \right\|^2 + |E_t^{\perp}|. \tag{5}$$

The detailed derivations are provided in Appendix B.2. Eq. (4) expresses the stepwise score shift as selected first-order terms plus a remainder. We use the $B^f_{\ell,i}$ coordinates to isolate the direct source-indexed role span attention pathway, while effects outside these coordinates remain in R_t. We define the selected first-order score contribution associated with (ℓ, i) as

$$C^{B,f}_{\ell,i,t} = \left\langle \nabla_{B^f_{\ell,i}} \psi_t, \Delta B^f_{\ell,i,t}(y) \right\rangle. \tag{6}$$

Note that $C^{B,f}_{\ell,i,t}$ is an inner product between the local score gradient and the matched contribution difference, so it gives a directional contribution to the score shift. Hence, Eq. (6) makes two factors explicit: (1) the matched difference $\Delta B^f_{\ell,i,t}(y)$ captures the role-dependent change in the forward-pass quantity, and (2) the gradient $\nabla_{B^f_{\ell,i}} \psi_t$ captures the local sensitivity of the stepwise score to this quantity. Equivalently, writing the inner product in norm-angle form gives

$$C^{B,f}_{\ell,i,t} = \|\nabla_{B^f_{\ell,i}} \psi_t\| \, \|\Delta B^f_{\ell,i,t}(y)\| \, \cos\angle \left(\nabla_{B^f_{\ell,i}} \psi_t, \Delta B^f_{\ell,i,t}(y) \right). \tag{7}$$

To this end, we identify three distinct factors: the matched difference $\Delta B^f_{\ell,i,t}(y)$, the local score sensitivity, and their directional alignment. Summing Eq. (4) over response positions and using Eq. (6) gives

$$\Phi_f(y) = \sum_{t=1}^{T} \sum_{(\ell,i)\in\mathcal{B}_{f,t}} C^{B,f}_{\ell,i,t} + R, \quad R = \sum_{t=1}^{T} R_t. \tag{8}$$

Together, Eqs. (7) and (8) show how local role span contributions organize into a trajectory-level score shift. Locally, the contribution magnitude depends on both the matched change in $B^f_{\ell,i}$ and the sensitivity of the stepwise score, while the directional alignment determines whether the term supports or opposes the score shift. Across the generation, the signed terms must aggregate coherently rather than cancel. The first-order decomposition therefore identifies two conditions for a trajectory-level effect, namely sufficient magnitude and coherent signed aggregation.

5 Experiments

In this section, we investigate the mechanism underlying the common dependence on context and the distinct generation outcomes of ICL and OOC hallucination. We use rule-following tasks derived from CL-bench (Dou et al., 2026) and evidence-grounding examples from RAGTruth (Niu et al., 2024), with settings provided in Appendix C. Our experiments examine three claims: (i) that role span contributions shift response preference in both settings, (ii) that rule and evidence realize this influence through different types of key token decisions, and (iii) that role-conditioned generation requires coherent contribution organization and sustained support at the relevant decisions. Appendices D.4 and D.5 extend the analysis to broader ICL and evidence-grounding settings.

5.1 Contribution Organization and Response Preference

To examine the relation between role span contribution and response preference, we construct matched candidate pairs for each role. Each pair contains a role-compatible response and a matched alternative that violates the role context. We define $M_f = \Phi_f(y^+) - \Phi_f(y^-)$ as the difference in response trajectory score shift between the two candidates, with $M_f > 0$ indicating that the contribution favors the role-compatible response over its matched alternative. Construction details are provided in Appendix C.3.

As shown in Figure 3(a), we empirically observe that $\mathrm{frac}(M_f > 0)$ is high for both roles, indicating that the role span contribution consistently favors the response compatible with the current role, rather than its matched alternative. More specifically, this preference is expressed as a tendency toward rule-following responses in ICL and evidence-grounded responses in the OOC setting.

Since the contribution propagates through the forward pass, its final effect accumulates across layers. The decomposition in Theorem 1 suggests two conditions for a substantial first-order contribution: (1) the contribution magnitude must be sufficiently large, and (2) the signed layerwise terms must aggregate coherently. Based on Eq. (7), we use $\|\nabla\psi_t\| \cdot \|\Delta B_{\ell,t}\|$ to measure contribution magnitude and the cosine term to determine the signed direction. Note that the aggregated first-order terms can track the exact score shift (Appendix D.2).

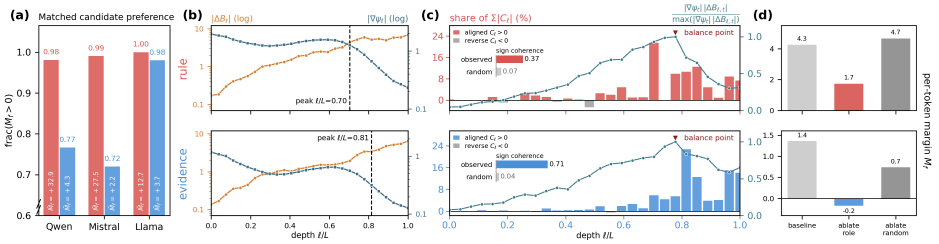


Figure 3: **Response preference and contribution organization**. **(a)** Fraction of positive candidate margins. For Qwen2.5-7B, **(b)** layerwise profiles of $|\Delta B_{\ell,t}|$ and $|\nabla\psi_t|$, **(c)** signed layerwise contributions and sign coherence, and **(d)** preference margin M_f under baseline, role span pathway ablation, and random ablation. More results are in Appendix D.1.

We first examine the layerwise contribution magnitude. As illustrated in Figure 3(c), both rule and evidence exhibit a localized middle-to-late concentration with a clear interior peak, indicating that substantial contribution is confined to a limited depth range. To understand the mechanism behind this peak, we further decompose the contribution magnitude into its two factors. As shown in Figure 3(b), $|\Delta B_{\ell,t}|$ and $|\nabla\psi_t|$ exhibit opposing trends: $|\Delta B_{\ell,t}|$ increases with depth, whereas $|\nabla\psi_t|$ decays. Given these opposing trends, their product is expected to peak at an intermediate depth where the increase in $|\Delta B_{\ell,t}|$ no longer outweighs the decrease in $|\nabla\psi_t|$. Consistent with this prediction, the resulting contribution exhibits the localized interior peak shown in Figure 3(c). In early layers, $\|\Delta B_{\ell,t}\|$ remains close to zero, indicating that the role span has induced a limited change. The contribution therefore remains small despite high downstream sensitivity. Thus, a high contribution appears only when both factors remain sufficiently large.

Notably, magnitude alone is insufficient if the signed terms cancel after aggregation. We further compare targeted contribution ablation with a random control, which controls for generic intervention effects. As shown in Figure 3(d), targeted ablation produces a larger reduction in M_f for both roles, showing that disrupting the contribution organization markedly weakens the resulting preference shift. Taken together, these results show that substantial contribution emerges when both magnitude factors remain sufficiently large, while disrupting their signed organization weakens the resulting preference shift.

5.2 Role Span Influence at Key Token Decisions

The first-order decomposition applies at each decoding step, allowing us to examine how role span influence varies across output. Intuitively, a role span should primarily affect generation through several key token choices. We test this by measuring the token-level first-order contribution $\hat{\delta}_t$ at the corresponding output decisions. As shown in Table 2(a), these decisions consistently occur at positions with high role span influence. We then remove the role span pathway and record the first change in token selection during free generation. Moreover, the first divergence after pathway removal likewise occurs at positions with high role span influence. Together, these findings support the insight that **role span shapes generation through key token choices.**

We further investigate the content of the divergence decisions. Using three generation conditions in Table 2(b), we examine whether role information is retained in the output. Under C_f, the required content is largely retained, confirming that the decisions are realized in the original generation. After pathway removal, retention drops toward the counterfactual level, indicating that the divergence decisions lose the role information. This comparison reveals a clear role-dependent pattern. Removing the rule pathway disrupts the specified format and reduces retention of the selected entity, whereas removing the evidence pathway changes the facts stated in the response. This pattern further clarifies the distinct functions of the two roles. Rule span contributions are typically realized

Table 2: **Role span effects on output decisions.** **(a)** Divergence rate and role span influence at the first divergence. **(b)** Retention rate of the content specified by C_f under the original, pathway-removed, and counterfactual conditions. Additional results are provided in Appendix D.6.

	(a) First-divergence localization (%)						(b) Retention of role-specified output content (%)								
	Rule span			Evidence span			Rule: format structure			Rule: selected entity			Evidence: stated fact		
Model	Div.	Attr.	Dec.	Div.	Attr.	Dec.	C_f	Removed	$\widetilde{C}_f$	C_f	Removed	$\widetilde{C}_f$	C_f	Removed	$\widetilde{C}_f$
Qwen2.5-7B	100.0	87.5	85.0	100.0	75.0	92.5	100.0	4.8	0.0	57.1	18.1	2.9	100.0	10.2	10.2
Mistral-7B	100.0	82.5	85.7	100.0	87.5	77.5	100.0	2.4	0.0	58.1	33.3	1.0	100.0	16.7	16.7
Llama-3.1-8B	100.0	80.0	90.0	100.0	85.0	95.0	100.0	4.8	0.0	53.3	18.1	7.6	100.0	10.0	13.3

Note. *Div.* reports the fraction of generations with changed token selection after pathway removal. *Attr.* and *Dec.* report role span influence at the first divergence and output decision positions, respectively; higher values indicate stronger influence.

through structural and selection decisions, whereas evidence span contributions are realized through factual content decisions. By shaping different types of key token decisions, the two roles give rise to distinct failure patterns in ICL and OOC hallucination.

5.3 Role Span Contributions to LLM Decision-Making

Building on the key token choices identified above, we now examine how role span influence becomes effective at decision level. We argue that the role span contribution should sustain support when LLM makes these key choices. To validate this, we first attenuate the signed contribution at initially successful decisions (Section 5.3.1) and then restore the contribution at naturally failing ones (Section 5.3.2).

5.3.1 Dependence of Successful Decisions

We first examine the structural and selection choices governed by the rule. At these choices, the role span contribution sustains the score advantage of the rule-specified token. We measure this preference by the gap between its logit and the highest remaining logit, with zero marking the decision boundary. Notably, attenuating the contribution at a single choice does not always alter the final response, as the role span remains visible to later decoding under causal attention. We therefore apply the attenuation throughout the continuation.

As shown in Table 3(a), increasing α generally narrows the margin under both roles. For rule decisions, the median margin shifts from positive to negative across all models, leaving the rule-specified token selected in only a minority of these key choices. Attenuation substantially erodes the advantage of the evidence-supported token, bringing the corresponding factual choice closer to the decision boundary. In contrast, matched attenuation of non-role spans produces smaller shifts, showing that the pronounced changes are driven by role span contributions. Under full attenuation, these margin reductions are accompanied by frequent token flips and performance declines on both the rule-following and evidence-grounded tasks, as shown in Table 3(b).

Table 3: **Sustained attenuation of role span contributions**. The contribution is scaled by $(1-\alpha)$, with α=1 denoting complete attenuation. **(a)** Median target token margin under targeted attenuation and matched non-role controls. **(b)** Token flips and task performance under full attenuation.

	(a) Target token margin							(b) Token and task outcomes at α=1			
	Targeted attenuation					Matched control		Targeted attenuation		Matched control	
Model	α=0	α=0.25	α=0.5	α=0.75	α=1	α=0	α=1	Flip (%)	Task (%)	Flip (%)	Task (%)
Rule span											
Qwen2.5-7B	4.62	5.00	-0.38	-12.89	-16.25	4.62	5.00	84.0	22.3	17.0	100.0
Mistral-7B	6.69	6.06	2.09	-4.16	-5.03	6.69	5.25	77.5	32.4	32.4	94.1
Llama-3.1-8B	4.69	3.25	2.06	-1.94	-4.06	4.69	2.38	76.5	28.6	25.5	86.7
Evidence span											
Qwen2.5-7B	8.23	7.90	5.63	-0.57	-1.09	8.23	7.60	57.4	25.9	9.3	74.1
Mistral-7B	10.00	9.94	6.72	2.56	1.25	10.00	9.88	44.2	42.3	0.0	98.1
Llama-3.1-8B	8.62	7.62	4.50	0.50	-0.69	8.62	7.75	54.9	37.3	2.0	98.0

Note. Let $y_t^\star$ denote the target token at this position, e.g., the token specified by the rule for rule decisions. We measure its local decision advantage using the *target token margin* $m_t = \log p_\theta(y_t^\star \mid P_t) - \max_{v \neq y_t^\star} \log p_\theta(v \mid P_t)$. Here, $m_t > 0$ indicates that the target token is top-ranked, while $m_t = 0$ marks the decision boundary.

These results show that the presence of role information does not guarantee effective support during generation. **The contribution must remain sufficiently large when LLM makes key choices.** For rules, this support must be sustained across repeated structural and selection decisions, whereas for evidence, it is required when the factual value is determined.

5.3.2 Recoverability of Failed Decisions

So far, we have focused on initially successful decisions, where attenuation shows that the local choice depends on sustained role span support. To test the causal effect of this support, we amplify the same contribution at naturally failing decisions. We track how far the target-token margin moves toward reversal and whether this local recovery improves task success.

As shown in Table 4, amplifying the evidence span contribution restores a substantial subset of naturally failing factual decisions. At $\beta = 1$, the target token is restored in 16.7%–34.3% across models, and the pooled restoration rate reaches 43.7% at $\beta = 2$, with task success improving as the contribution is amplified. While evidence failures exhibit substantial recovery, rule failures reveal a clear limit. Their initial margins lie much farther below the decision boundary, leaving a deficit too large for the target token to be restored even under $16\times$ amplification, while task success remains low. These results show that the same local contribution mechanism can extend to naturally failing decisions. Whether amplification reverses the choice depends on the balance between the support of role contribution and the decision deficit. This relation further yields a quantitative prediction for failed decisions, which we formalize in Appendix B.4 and evaluate in Appendix D.7.

Table 4: **Restoration at failing decisions**. We restore the contribution vector at the decision, scaled by a weight β. **(a)** Restoring the contribution at $\beta = 1$. **(b)** Restoration across multiple weights.

(a) Restoration at $\beta = 1$

Model	Baseline margin	Margin change	Selection (%)	Task (%)
Rule span				
Qwen2.5-7B	-14.6	-0.61	0.0	0.0
Mistral-7B	-19.0	+0.06	0.0	4.8
Llama-3.1-8B	-11.3	+0.00	0.0	0.0
Evidence span				
Qwen2.5-7B	-4.0	+0.89	29.5	34.1
Mistral-7B	-2.2	+0.28	16.7	20.8
Llama-3.1-8B	-2.7	+1.69	34.3	37.1

(b) Restoration across weights

Metric	$\beta = 0$	$\beta = 0.5$	$\beta = 1$	$\beta = 2$	$\beta = 4$	$\beta = 8$	$\beta = 16$
Selection (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Margin (median)	-14.6	-14.6	-14.5	-14.1	-13.0	-12.8	-11.7
Task (%)	0.0	0.0	1.3	2.5	6.3	3.8	6.3
Selection (%)	0.0	22.3	28.2	43.7	41.7	39.8	38.8
Margin (median)	-3.0	-2.9	-2.0	-1.2	-0.6	-1.9	-2.8
Task (%)	0.0	25.2	32.0	47.6	53.4	55.3	58.3

Note. Margins follow the definition in Table 3. The margin change is $\Delta m_t = m_t(\beta = 1) - m_t(\beta = 0)$. Positive values indicate amplification shifts the local decision toward the target token, whereas negative values indicate the opposite.

In summary, we show that dynamic context guides LLM generation through functional role span contributions in both ICL and OOC hallucination. Specifically, rule contributions shape structural and selection choices, whereas evidence contributions shape factual choices. Moreover, effective role influence depends on coherent contribution organization across layers and sufficient support when the key choice is made. The causal interventions further show that a failed choice can be reversed when the corresponding support is sufficient to compensate for the decision deficit.

6 Conclusion

In this paper, we offer a common mechanistic view of the distinct outcomes of ICL and OOC hallucination. This view treats the functional roles of dynamic context as the organizing principle for model behavior. We connect internal role representations to their effects on decoding through role span contributions. These contributions causally shape key token choices that determine response structure, content selection, and factual value. Each functional role determines where these contributions must remain effective during generation. We derive a first-order decomposition of role span contributions and identify two necessary conditions for reliable influence. At the decision level, rule following requires sustained support across repeated structural and selection choices, whereas evidence grounding relies on support at a smaller set of decisive factual choices. The contrasting outcomes therefore follow from distinct requirements for sustaining role support across generation.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks requiring mandatory disclosure under the ICLR 2027 AI Policy. Additionally, we used generative AI tools to create or edit software code and to edit the research paper to improve readability. We have reviewed all AI-assisted work. AI-assisted code was verified and tested by the authors, and AI-assisted edits to the manuscript were reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The source code used for our experiments is provided in the supplementary materials and will be released in a public repository. Appendix B provides the complete proofs and derivations of the theoretical results. Appendix C provides the experimental details, including the experimental setting, matched role construction, representation and recovery analyses, and contribution and preference analysis. Appendix D reports additional experimental results.

REFERENCES

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188, 2024.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Hakaze Cho, Mariko Kato, Yoshihiro Sakai, and Naoya Inoue. Revisiting in-context learning inference circuit in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 13081–13117, 2025.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, et al. Cl-bench: A benchmark for context learning. *arXiv preprint arXiv:2602.03587*, 2026.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. Context versus prior knowledge in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13211–13235, 2024.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Pre-training to learn in context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4849–4870, 2023.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yangfan Ye, Weihong Zhong, Yuxuan Gu, Baoxin Wang, Dayong Wu, et al. Improving contextual faithfulness of large language models via retrieval heads-induced optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16896–16913, 2025a.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM transactions on information systems*, 43(2):1–55, 2025b.

Xin Li and Dan Roth. Learning question classifiers. In *Coling 2002: The 19th international conference on computational linguistics*, 2002.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022.

Clement Neo, Shay B Cohen, and Fazl Barez. Interpreting context look-ups in transformers: Investigating attention-mlp interactions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16681–16697, 2024.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9090–9101, 2024.

Jie Ren, Qipeng Guo, Hang Yan, Dongrui Liu, Quanshi Zhang, Xipeng Qiu, and Dahua Lin. Identifying semantic induction heads to understand in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6916–6932, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International conference on machine learning*, pp. 31210–31227. PMLR, 2023.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Xubin Wang, Jianfei Wu, Yichen Yuan, Deyu Cai, Mingzhe Li, and Weijia Jia. Demonstration selection for in-context learning via reinforcement learning. *arXiv preprint arXiv:2412.03966*, 2024.

Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Advances in neural information processing systems*, 37: 33402–33422, 2024.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, volume 2025, pp. 62143–62156, 2025.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, volume 2024, pp. 35623–35646, 2024.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pp. 39818–39833. PMLR, 2023.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, volume 2024, pp. 29862–29883, 2024.

Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. Mechanistic understanding and mitigation of language model non-factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7943–7956, 2024.

Appendix

A. Related Work

Context-conditioned generation. Motivated by the contextual adaptability of LLMs, recent work has extensively studied how information supplied at inference time shapes model behavior. In-context learning provides a prominent setting where models can infer a task from demonstrations and instructions without task-specific training (Gu et al., 2023; Ye et al., 2023). Recent studies further show that this adaptation is sensitive to demonstration selection, ordering, and composition (Lu et al., 2022; Ye et al., 2023; Peng et al., 2024; Wang et al., 2024). Meanwhile, a separate line of work has investigated hallucination in LLM generation, where model outputs contain factually incorrect or unsupported information (Huang et al., 2025b; Yu et al., 2024). Retrieval-augmented generation has become a common approach to mitigating hallucination by incorporating external evidence into generation (Yoran et al., 2024; Asai et al., 2024). Nevertheless, LLMs may still produce claims that are unsupported by or contradictory to the retrieved content (Niu et al., 2024; Wu et al., 2024). Such departures from the supplied context define the out-of-context (OOC) hallucination setting and expose a gap in understanding contextual faithfulness (Asai et al., 2024; Huang et al., 2025a).

Mechanistic analyses of inference-time behavior. Recent mechanistic work extends the study of inference-time behavior from observable outcomes to internal representations. In ICL, some studies identify attention circuits that support in-context computation (Ren et al., 2024), while other work seeks to explicitly characterize in-context task information by recovering task vectors from hidden states (Hendel et al., 2023; Todd et al., 2024). In parallel, mechanistic studies of hallucination associate incorrect or unsupported generation with failures in knowledge retrieval and transfer and with conflicts between parametric knowledge and external evidence (Yu et al., 2024; Xie et al., 2024). Other studies further localize components involved in retrieving and using contextual information during generation, showing that evidence use can likewise be traced to internal components (Wu et al., 2025; Neo et al., 2024). Despite this progress, the two behaviors are predominantly examined within specific tasks or failure settings. Across both lines, a common concern is which internal representations of context become consequential for generation.

Attributing contextual influence. To establish this relevance, existing methods connect model predictions to their internal sources. For instance, input attribution and score decomposition quantify how tokens or activations support a prediction (Sundararajan et al., 2017), while circuit analysis traces it to interpretable computational paths (Dunefsky et al., 2024). However, such local analyses rarely make explicit the function of context in the current task (Yu et al., 2024). Despite comparable mechanistic accounts, ICL and OOC hallucination exhibit distinct task-level behaviors. To this end, we analyze the distinct functional roles in the two settings and identify the mechanistic conditions underlying task success and failure from a unified perspective.

B. Proofs and Derivations

B.1. Attention-Mediated Role Span Transport

We first formalize why a role-induced difference can enter a new token position only through attention.

Lemma 1 (Position locality). *In a decoder-only Transformer, normalization and MLP modules act independently at each token position, whereas the attention update at position i aggregates states from the source positions visible to i.*

Proof. Normalization and MLP modules apply the same map independently at each position. By contrast, Eq. (2) writes the attention update at i as a sum over its visible source positions. Position-wise transformations of queries and keys do not alter this distinction. □

Fix a decoding step t and write $P = P_t^f(y)$, $\widetilde{P} = \widetilde{P}_t^f(y)$, and $\mathcal{S}_f$ for their aligned role span positions. Define

$$d_\ell(i) = h_\ell(P, i) - h_\ell(\widetilde{P}, i), \qquad \mathcal{C}_\ell = \{s : d_\ell(s) \neq 0\},$$

and let $\mathcal{V}(i)$ denote the source positions visible to i. For head a and source position $s \in \mathcal{V}(i)$, write

$$g_{\ell,i,s}^{(a)}(P) = \alpha_{i\leftarrow s}^{(a)}(P) V_\ell^{(a)}(P, s) W_O^{(a)}, \qquad \Delta g_{\ell,i,s}^{(a)} = g_{\ell,i,s}^{(a)}(P) - g_{\ell,i,s}^{(a)}(\widetilde{P}).$$

We decompose the source-indexed differences as

$$\Delta B_{\ell,i}^f = \sum_a \sum_{s \in \mathcal{V}(i) \cap \mathcal{S}_f} \Delta g_{\ell,i,s}^{(a)},$$

$$\Delta I_{\ell,i} = \sum_a \sum_{s \in \mathcal{V}(i) \cap (\mathcal{C}_{\ell-1} \setminus \mathcal{S}_f)} \Delta g_{\ell,i,s}^{(a)},$$

$$\Delta W_{\ell,i} = \sum_a \sum_{s \in \mathcal{V}(i) \setminus (\mathcal{S}_f \cup \mathcal{C}_{\ell-1})} \Delta g_{\ell,i,s}^{(a)}.$$

Proposition 1 (Attention-mediated role span transport). *Suppose that P and $\widetilde{P}$ differ only at the aligned role span positions $\mathcal{S}_f$. For every $i \notin \mathcal{S}_f$, $d_0(i) = 0$, and at each layer $\ell \geq 1$,*

$$\Delta \operatorname{Attn}_\ell(i) = \Delta B_{\ell,i}^f + \Delta I_{\ell,i} + \Delta W_{\ell,i}, \tag{9}$$

where

$$\Delta \operatorname{Attn}_\ell(i) = \operatorname{Attn}_\ell^{\text{out}}(P, i) - \operatorname{Attn}_\ell^{\text{out}}(\widetilde{P}, i).$$

Moreover,

$$d_{\ell-1}(i) = 0 \quad \textit{and} \quad \Delta \operatorname{Attn}_\ell(i) = 0 \quad \Longrightarrow \quad d_\ell(i) = 0.$$

Consequently, if $\ell^\star = \min\{\ell : d_\ell(i) \neq 0\}$ exists, then $\Delta \operatorname{Attn}_{\ell^\star}(i) \neq 0$.

Proof. Since the matched inputs differ only at $\mathcal{S}_f$, their embedding representations coincide at every $i \notin \mathcal{S}_f$. Hence, $d_0(i) = 0$.

We next decompose the attention difference. By Eq. (2),

$$\Delta \operatorname{Attn}_\ell(i) = \sum_a \sum_{s \in \mathcal{V}(i)} \Delta g_{\ell,i,s}^{(a)}.$$

The visible source positions admit the disjoint partition

$$\mathcal{V}(i) = \big(\mathcal{V}(i) \cap \mathcal{S}_f\big) \,\dot\cup\, \big(\mathcal{V}(i) \cap (\mathcal{C}_{\ell-1} \setminus \mathcal{S}_f)\big) \,\dot\cup\, \big(\mathcal{V}(i) \setminus (\mathcal{S}_f \cup \mathcal{C}_{\ell-1})\big).$$

Summing over these three sets gives Eq. (9). The first term is the direct role span contribution defined in Eq. (3); the second reads from non-role positions whose residual states already differ.

For a source position in the third set, $d_{\ell-1}(s) = 0$, and hence its value representation is identical under the two computations:

$$V_\ell^{(a)}(P, s) = V_\ell^{(a)}(\widetilde{P}, s).$$

Thus, $\Delta W_{\ell,i}$ contains the remaining changes in attention weights over otherwise unchanged source values. This definition covers any such reweighting without assigning it to a particular query, key, or softmax component.

We next show how the difference propagates through the block. If $d_{\ell-1}(i) = 0$ and $\Delta \operatorname{Attn}_\ell(i) = 0$, the two computations provide the same local state and the same attention update at position i. By Lemma 1, all remaining operations at that position therefore receive identical inputs, which gives $d_\ell(i) = 0$.

Now let $\ell^\star$ be the first layer such that $d_{\ell^\star}(i) \neq 0$. Then $d_{\ell^\star-1}(i) = 0$. If the attention updates also agreed at layer $\ell^\star$, the previous implication would yield $d_{\ell^\star}(i) = 0$, a contradiction. Therefore, $\Delta \operatorname{Attn}_{\ell^\star}(i) \neq 0$. □

Proposition 1 establishes that attention is the direct channel through which a role-induced difference enters a new token position. Among the resulting terms, $\Delta B^f_{\ell,i}$ is the direct source-indexed role span pathway used in Section 4.3, while $\Delta I_{\ell,i}$ and $\Delta W_{\ell,i}$ are treated in the remainder analysis of Appendix B.3.

B.2. First-Order Role Span Score Decomposition

We now make the coordinate dependence in Assumption 1 explicit. Fix a decoding step t and write

$$\widetilde{b} = \Big(B^f_{\ell,i}(\widetilde{P}^f_t(y))\Big)_{(\ell,i)\in\mathcal{B}_{f,t}}, \quad b^f = \Big(B^f_{\ell,i}(P^f_t(y))\Big)_{(\ell,i)\in\mathcal{B}_{f,t}},$$

with

$$\Delta b = b^f - \widetilde{b} = \Delta B^f_{\mathcal{B},t}(y).$$

We equip this stacked coordinate space with the Euclidean inner product and norm.

Substitution map. For $b = (b_{\ell,i})_{(\ell,i)\in\mathcal{B}_{f,t}}$, let $\Psi_t(b)$ denote the score obtained from $\widetilde{P}^f_t(y)$ by processing the model in forward order and replacing each selected role span attention contribution with $b_{\ell,i}$ before it enters the residual stream. Specifically, the term recomputed from the current modified state is removed and replaced by $b_{\ell,i}$. All replacements within a layer are applied before advancing to the next layer, and the remaining quantities are computed from the resulting states.

Lemma 2 (Exact substitution identity). *The substitution map is well defined and satisfies*

$$\Psi_t(\widetilde{b}) = \psi_t(\widetilde{P}^f_t(y)).$$

Consequently, with

$$E^{\perp}_t = \psi_t(P^f_t(y)) - \Psi_t(b^f),$$

the stepwise score shift admits the exact identity

$$\delta^f_t(y) = \Psi_t(b^f) - \Psi_t(\widetilde{b}) + E^{\perp}_t. \tag{10}$$

Proof. The forward ordering uniquely determines every modified residual state, so $\Psi_t(b)$ is well defined for any collection b.

We next establish the endpoint identity. Set $b = \widetilde{b}$. At the input layer, the modified computation and the computation under $\widetilde{P}^f_t(y)$ have identical states. Suppose that their states agree through layer $\ell - 1$. Their unmodified attention terms at layer ℓ are then identical, and each selected role span term is replaced by its value $B^f_{\ell,i}(\widetilde{P}^f_t(y))$. Hence, the complete attention updates and the resulting residual states also agree at layer ℓ. Induction over the layers gives $\Psi_t(\widetilde{b}) = \psi_t(\widetilde{P}^f_t(y))$.

Finally, adding and subtracting $\Psi_t(b^f)$ in the definition of $\delta^f_t(y)$ gives

$$\begin{aligned}\delta^f_t(y) &= \psi_t(P^f_t(y)) - \psi_t(\widetilde{P}^f_t(y)) \\ &= \Psi_t(b^f) - \Psi_t(\widetilde{b}) + \big[\psi_t(P^f_t(y)) - \Psi_t(b^f)\big],\end{aligned}$$

which is Eq. (10). □

Proof of Theorem 1. By Lemma 2, it remains to expand $\Psi_t(b^f) - \Psi_t(\widetilde{b})$. Let

$$\gamma(\tau) = \widetilde{b} + \tau\Delta b, \qquad \tau \in [0,1].$$

Then we have

$$\Psi_t(b^f) - \Psi_t(\widetilde{b}) = \int_0^1 \langle \nabla\Psi_t(\gamma(\tau)), \Delta b\rangle \, d\tau.$$

Adding and subtracting $\nabla\Psi_t(\widetilde{b})$ gives

$$\Psi_t(b^f) - \Psi_t(\widetilde{b}) = \Big\langle \nabla\Psi_t(\widetilde{b}), \Delta b\Big\rangle + r^{\mathrm{curv}}_t,$$

where

$$r_t^{\mathrm{curv}} = \int_0^1 \left\langle \nabla\Psi_t(\gamma(\tau)) - \nabla\Psi_t(\widetilde{b}), \Delta b \right\rangle d\tau.$$

Since Ψ_t has an L_t-Lipschitz gradient on the segment between $\widetilde{b}$ and b^f,

$$\begin{aligned} |r_t^{\mathrm{curv}}| &\le \int_0^1 \left\| \nabla\Psi_t(\gamma(\tau)) - \nabla\Psi_t(\widetilde{b}) \right\| \|\Delta b\| \, d\tau \\ &\le \int_0^1 L_t \tau \|\Delta b\|^2 \, d\tau = \frac{L_t}{2} \|\Delta b\|^2. \end{aligned}$$

Moreover, the stacked inner product decomposes as

$$\left\langle \nabla\Psi_t(\widetilde{b}), \Delta b \right\rangle = \sum_{(\ell,i)\in\mathcal{B}_{f,t}} \left\langle \nabla_{b_{\ell,i}} \Psi_t(\widetilde{b}), \Delta B^f_{\ell,i,t}(y) \right\rangle.$$

Combining this equality with Eq. (10) yields Eq. (4) with

$$R_t = r_t^{\mathrm{curv}} + E_t^{\perp}.$$

Finally,

$$|R_t| \le |r_t^{\mathrm{curv}}| + |E_t^{\perp}| \le \frac{L_t}{2} \left\| \Delta B^f_{\mathcal{B},t}(y) \right\|^2 + |E_t^{\perp}|,$$

which gives Eq. (5). □

The term r_t^{curv} is the Taylor remainder within the selected coordinate space, whereas $E_t^{\perp}$ contains the score change not reproduced by these coordinates. We characterize the latter terms in Appendix B.3.

B.3. Remainder Characterization

By Theorem 1, the remainder separates as $R_t = r_t^{\mathrm{curv}} + E_t^{\perp}$, where $|R_t| \le \frac{L_t}{2} \left\| \Delta B^f_{\mathcal{B},t}(y) \right\|^2 + |E_t^{\perp}|$.

Here, r_t^{curv} is the curvature error introduced by linearizing Ψ_t over the selected coordinates, whereas $E_t^{\perp} = \psi_t(P_t^f(y)) - \Psi_t(b^f)$ measures the score change not reproduced by these substitutions. Relative to the decomposition in Appendix B.1, the substitution gap may include omitted direct coordinates and portions of $\Delta I_{\ell,i}$ and $\Delta W_{\ell,i}$ that are not recovered by the modified computation.

Notably, this characterization does not assume that R_t is small. In Appendix D.3, we compare the direct role span attention pathway with alternative transport terms and MLP updates, and show that attention provides the dominant route for role information to reach the relevant token decisions.

B.4. Decision-Level Recovery Prediction

We finally relate the role span contribution to the local decision boundary considered in Section 5.3. For a naturally failing decision, let $y_t^\star$ denote the target token and let

$$c_t = \arg\max_{v \neq y_t^\star} \log p_\theta(v \mid P_t; \beta = 0)$$

be its strongest competitor before amplification. Following Section 5.3.2, let β parameterize amplification of the same role span contribution, with $\beta = 0$ denoting the original decision. For the fixed competitor c_t, write

$$m_t(\beta; c_t) = \log p_\theta(y_t^\star \mid P_t; \beta) - \log p_\theta(c_t \mid P_t; \beta),$$

so that $m_t(0; c_t) = m_t(0) < 0$. We define the signed support supplied along the amplification direction as

$$S_t = \frac{d}{d\beta} m_t(\beta; c_t) \Big|_{\beta=0}.$$

Proposition 2 (Decision-boundary crossing). *Suppose that the derivative of $m_t(\beta; c_t)$ is L-Lipschitz along the local amplification path. Then*

$$|m_t(\beta; c_t) - m_t(0) - \beta S_t| \leq \frac{L}{2}\beta^2. \tag{11}$$

For $m_t(0) < 0$ and $S_t > 0$, the corresponding first-order prediction of the amplification required to reach the decision boundary is

$$\widehat{\beta}_t = -\frac{m_t(0)}{S_t}. \tag{12}$$

Moreover, if c_t remains the strongest non-target token along the local path, then any $\beta > 0$ satisfying

$$\beta S_t > -m_t(0) + \frac{L}{2}\beta^2 \tag{13}$$

guarantees $m_t(\beta) > 0$.

Proof. A first-order expansion of the fixed-competitor margin at $\beta = 0$ gives

$$m_t(\beta; c_t) = m_t(0) + \beta S_t + \varepsilon_t(\beta),$$

where the Lipschitz condition yields $|\varepsilon_t(\beta)| \leq L\beta^2/2$. This proves Eq. (11).

Ignoring the quadratic term, the local boundary is reached when $m_t(0) + \beta S_t = 0$. For $S_t > 0$, solving this equality gives Eq. (12). Thus, a larger initial deficit $-m_t(0)$ requires stronger amplification, whereas larger positive signed support S_t reduces the required strength.

For the finite-β statement, Eq. (11) gives the lower bound

$$m_t(\beta; c_t) \geq m_t(0) + \beta S_t - \frac{L}{2}\beta^2.$$

Hence, Eq. (13) implies $m_t(\beta; c_t) > 0$. If c_t remains the strongest competitor, this pairwise margin equals the target-token margin, and therefore $m_t(\beta) > 0$. □

Consequently, the first-order decision formulation estimates the amplification required to recover a failed decision. Appendix D.7 further examines the extension of this prediction to task-level recovery.

C. Experimental Details

Experimental setting. For the rule role, we derive a controlled benchmark from the task patterns in CL-bench (Dou et al., 2026), covering format, selection, and include or omit instructions. For the evidence role, we use the training split of RAGTruth (Niu et al., 2024). We evaluate Qwen2.5-7B, Mistral-7B, and Llama-3.1-8B. Free generations use greedy decoding, whereas fixed response trajectories are evaluated with teacher forcing.

C.1. Matched Role Construction

To isolate functional role f, we compare two prompts that share the query and all non-role content but differ in the role-bearing span. The matched counterfactual changes only this span while preserving the remainder of the prompt. Figure 4 illustrates the matched construction. To ensure alignment, we locate the edited span separately in each tokenized prompt and, for a fixed response trajectory y, use the same response prefix $y_{<t}$ under both conditions, with comparisons performed at the same decoding step.

C.2. Representation and Recovery Analyses

Subspace concentration. We extract Δh_ℓ^f at the final prompt position from every decoder layer. For each model and role, we fit a rank-10 subspace on one half of the samples and evaluate it on the other, then exchange the two splits. We use the matched role comparisons defined in Appendix C.1.

Rule role	Matched counterfactual
C_r**: role-bearing span**	$\widetilde{C}_r$**: counterfactual span**
Rule: When summarising the records below, output the answer as a JSON object with keys `name` and `value`.	Rule: When summarising the records below, output the answer as a Markdown bullet list, one bullet per record.
Shared: records, task description, query, and remaining prompt.	

Evidence role	Matched counterfactual
C_e**: role-bearing span**	$\widetilde{C}_e$**: counterfactual span**
Evidence: Bake in the hot oven for **20** minutes.	Evidence: Bake in the hot oven for **30** minutes.
Shared: remaining document, grounding instruction, query, and prompt template.	

Figure 4: Simple examples of matched role comparisons.

Specifically, the held-out concentration is computed as $\mathcal{E}^f_{\ell,i} = \frac{\|\Pi_{U_\ell}\Delta h^f_\ell(i)\|_2^2}{\|\Delta h^f_\ell(i)\|_2^2}$, where Π_{U_ℓ} projects onto the fitted rank-10 subspace U_ℓ. The random-subspace baseline is $\mathcal{E}_{\text{rand}} = 10/H$, where H is the hidden state dimension.

Output alignment and behavior recovery. We assess output alignment by applying the final RMS normalization to the two residual states in each matched pair and then projecting their difference onto the unembedding direction of the token selected under C_f. We report the resulting logit shift and cosine alignment, with a matched non-role perturbation as the control. Separately, we evaluate behavior recovery by adding Δh^f_ℓ to the counterfactual state at the final prompt position and recording whether the next-token choice returns to that under C_f.

C.3. Contribution and Preference Analysis

Candidate comparisons. Each preference analysis combines a matched context pair with a minimal candidate pair. Both candidates are scored with teacher forcing, and M_f measures the resulting change in relative support. Figure 5 illustrates the design.

	Matched contexts	**Matched candidates**
Rule	C_r: JSON-output rule $\widetilde{C}_r$: Markdown-list rule	y^+: {"`name`": "Alpha", "value": "120"} y^-: {"`label`": "Alpha", "value": "120"}
Evidence	C_e: Evidence supporting `2024` $\widetilde{C}_e$: Matched context, e.g., evidence supporting `2029`	y^+: The value is `2024` y^-: The value is `2029`

$$\Phi_f(y) = \log p_\theta(y \mid C_f \oplus x) - \log p_\theta(y \mid \widetilde{C}_f \oplus x), \quad M_f = \Phi_f(y^+) - \Phi_f(y^-).$$

Figure 5: **Matched preference construction**. Each context pair differs only in the role-bearing span, while each candidate pair differs only in the choice specified by that role.

D. Additional Experiments

D.1. Additional Model Results for Contribution Organization

Figures 6 and 7 report the corresponding analyses for Mistral-7B and Llama-3.1-8B, both of which exhibit the same overall pattern. Contributions remain concentrated within a limited range of layers, the two magnitude factors vary in opposite directions across depth, and role span pathway ablation reduces M_f more than random ablation. The contribution organization identified in the main analysis is therefore consistent across models.

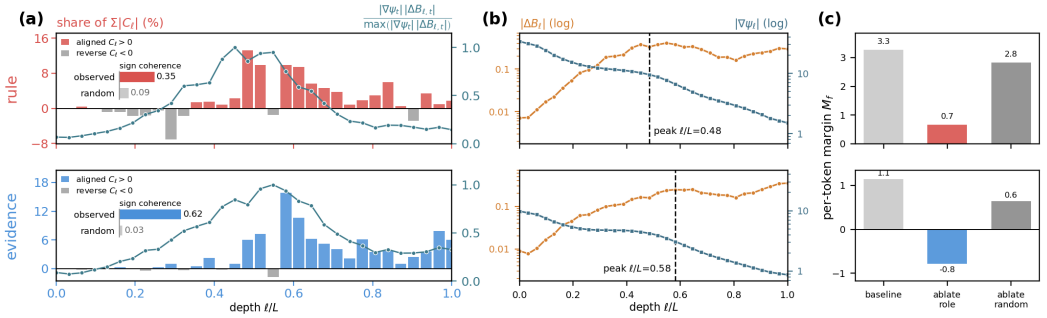


Figure 6: **Organization of contributions for Mistral-7B**. **(a)** Signed layerwise contributions and sign coherence; **(b)** layerwise profiles of $|\Delta B_{\ell,t}|$ and $|\nabla \psi_t|$; and **(c)** preference margin M_f under the baseline, role span pathway ablation, and random ablation.

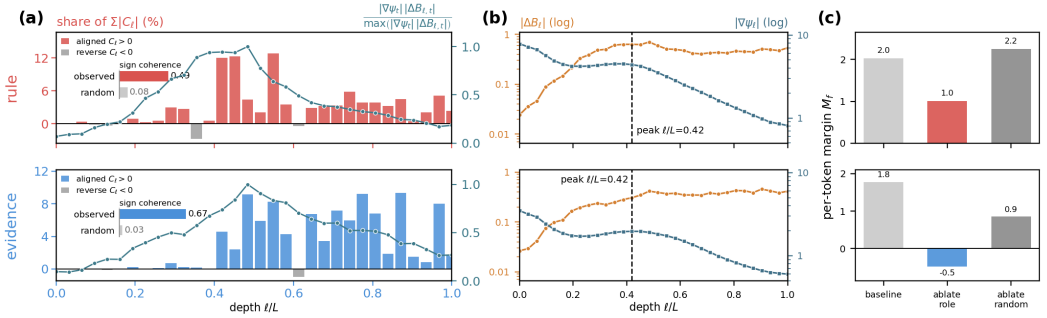


Figure 7: **Organization of contributions for Llama-3.1-8B**. **(a)** Signed layerwise contributions and sign coherence; **(b)** layerwise profiles of $|\Delta B_{\ell,t}|$ and $|\nabla \psi_t|$; and **(c)** preference margin M_f under the baseline, role span pathway ablation, and random ablation.

D.2. First-Order Correspondence

Here we examine how the first-order contributions correspond to the exact role effect. Within each response, the magnitudes of the summed first-order contributions track the exact stepwise score shifts, with median Spearman correlations of 0.85–0.94 across models. The sign of the accumulated contribution also matches the exact trajectory shift for most responses.

D.3. Role Span Attention Pathway

Appendix B.1 establishes that attention is the direct channel when a role-induced difference enters a new token position. To examine how much of the realized role effect this pathway carries, we study its recoverability across attention terms and its decision effects. Specifically, for each matched context pair, we fix the exact trajectory score effect Φ_f, and substitute one term at a time. We consider the direct role term ΔB, attention reweighting ΔW, and a matched non-role term.

Table 5: **Correspondence between first-order contributions and exact score shifts**.

	Rule span		Evidence span	
Model	Corr.	Sign (%)	Corr.	Sign (%)
Qwen2.5-7B	0.93	77.1	0.86	95.0
Mistral-7B	0.94	87.6	0.88	97.0
Llama-3.1-8B	0.92	91.4	0.85	94.0

Note. *Corr.* is the median Spearman correlation between the magnitudes of the summed first-order contributions and the exact stepwise score shifts. *Sign* is the fraction of responses whose accumulated contribution matches the sign of the exact trajectory shift.

Table 6 shows that the direct term recovers 73.5%–93.9% of the exact role effect across models. In contrast, recovery from attention reweighting and matched non-role positions remains near zero. Together with the sustained attenuation results in Section 5.3.1, these findings show that the direct role span attention term carries a substantial and behaviorally consequential part of the role effect.

Since it is essential for the role effect to become available at the decision, we next attenuate the carried state at three depths. Table 7(a) shows that attenuation at $0.25L$ has little effect in any setting. By $0.50L$, the dependence begins to emerge, and at $0.75L$, the margin drop is positive for all settings. Thus, the role effect becomes relevant to the decision only after sufficient depth. Table 7(b) shows that attenuating either update family reduces the target margin. Across comparisons, the attention intervention produces a larger effect than the MLP intervention.

Taken together, these results establish the empirical importance of the role span attention pathway. The direct attention term recovers a substantial part of the role effect, and this effect becomes increasingly available from the carried state with depth. Combined with the consistently larger intervention effect of attention than MLP, this pattern supports attention as the dominant route through which role information reaches the relevant token decisions.

Table 6: **Direct recovery of role effects.** Median normalized recovery (%) after substituting the direct role term (ΔB), attention reweighting (ΔW), or matched non-role positions.

	Rule			**Evidence**		
Model	Direct ΔB	Reweighting ΔW	Non-role	Direct ΔB	Reweighting ΔW	Non-role
Qwen2.5-7B	73.5	0.0	0.1	85.9	1.9	1.4
Mistral-7B	88.3	0.6	-0.9	92.0	1.9	-0.7
Llama-3.1-8B	87.7	0.8	-0.2	93.9	0.4	0.8

Table 7: **Accumulated carry and update-family effects at role-sensitive decisions. (a)** The carried state entering the layer at three fractional depths. **(b)** Each update family at all layers.

	(a) Carry drop by depth			**(b) Update families**	
Model	ℓ/L=0.25	0.5	0.75	Attention	MLP
Rule					
Qwen2.5-7B	-0.01 [-0.13, 0.12]	-0.22 [-0.64, 0.14]	11.25 [7.18, 16.11]	30.79 [24.51, 37.68]	26.14 [21.55, 31.06]
Mistral-7B	0.05 [-0.01, 0.11]	0.77 [0.15, 1.42]	12.56 [9.76, 15.32]	15.64 [12.20, 19.21]	13.40 [10.84, 15.97]
Llama-3.1-8B	-0.04 [-0.10, 0.03]	3.25 [2.10, 4.55]	9.11 [8.20, 10.00]	13.74 [11.59, 15.96]	8.77 [7.71, 9.68]
Evidence					
Qwen2.5-7B	0.02 [-0.24, 0.28]	0.23 [-0.11, 0.61]	4.25 [3.05, 5.59]	8.76 [6.76, 11.01]	7.18 [5.43, 9.02]
Mistral-7B	-0.04 [-0.17, 0.07]	1.10 [0.68, 1.53]	5.00 [3.83, 6.19]	8.70 [6.85, 10.83]	7.81 [6.37, 9.27]
Llama-3.1-8B	0.20 [-0.06, 0.41]	2.61 [1.82, 3.46]	4.68 [3.61, 5.76]	7.90 [6.22, 9.70]	4.01 [2.62, 5.54]

Note. Drops are computed as matched control minus targeted intervention, so larger values indicate stronger dependence.

D.4. Rule Contributions in Few-Shot ICL

Our main experiments place each rule in an explicit prompt span, allowing us to isolate its effect on generation. To test whether the same account extends to few-shot ICL, we use prompts in which labeled examples implicitly specify the rule. The model must infer the underlying task mapping from these examples. We then test whether the inferred mapping makes a measurable causal contribution to the decision and identify where this contribution is realized within the examples.

We use the SST-2 sentiment task from the Stanford Sentiment Treebank (Socher et al., 2013) and the six-class TREC question classification task (Li & Roth, 2002). To isolate the effect of the inferred mapping, we preserve the demonstrations, their order, the query, and the prompt format across each prompt pair, varying only the assignment of labels to classes. We use the demonstration blocks as the rule positions and matched positions outside them as controls.

Table 8(a) shows that all three models follow the induced mapping, while breaking the input–label pairing reduces mapping following with the same inputs and labels preserved in the prompt. This indicates that the rule is induced from relations established across the demonstrations. Table 8(b)

Table 8: **Role effects for rules induced from demonstrations.** The three column groups report rule induction, contribution alignment, and the effect of targeted attenuation.

	(a) Rule induction			(b) Contribution		(c) Decision effect	
Model	Task acc. (%) orig./perm.	Mapping follow (%)	No-demo/ broken (%)	Positive margin (%)	Recovery (%) demo/control	Margin drop [95% CI]	Accuracy drop [95% CI] (pp)
SST-2							
Qwen2.5-7B	77.1 / 86.7	84.8	46.7 / 54.3	81.9	99.4 / 0.1	2.27 [1.7, 2.8]	25.7 [15.2, 35.2]
Mistral-7B	82.9 / 82.9	82.9	49.5 / 48.6	81.9	99.8 / 0.0	1.24 [0.8, 1.7]	20.0 [7.6, 31.4]
Llama-3.1-8B	73.3 / 77.1	79.0	41.0 / 41.9	77.1	96.2 / −1.0	0.63 [0.3, 0.9]	15.2 [6.7, 23.8]
TREC							
Qwen2.5-7B	42.9 / 46.7	65.7	20.0 / 25.7	76.2	99.4 / −0.1	1.04 [0.4, 1.7]	8.6 [−1.9, 20.0]
Mistral-7B	52.4 / 47.6	66.7	14.3 / 21.9	74.3	98.5 / −0.5	1.37 [0.9, 1.9]	28.6 [18.1, 39.0]
Llama-3.1-8B	35.2 / 37.1	73.3	14.3 / 17.1	66.7	97.0 / −0.7	0.69 [0.2, 1.1]	18.1 [8.6, 27.6]

Note. Recovery is evaluated on examples with $|\Phi_r| \geq 0.5$. Matched positions outside the demonstration blocks form the control. Drop metrics are computed as control minus targeted attenuation, so larger values indicate stronger dependence on the selected positions.

shows that the inferred rule yields positive contribution margins and that substituting the demonstration positions recovers nearly the full effect, with negligible recovery from the matched control. Table 8(c) further shows that targeted attenuation reduces the label margin in all six settings and significantly lowers task accuracy in five, with the 95% CI. Together, these results show that the inferred mapping produces a measurable role span contribution that affects the query decision. We next examine where this contribution is realized within the demonstrations.

Table 9: **Position-wise organization of induced role effects.** Recovery and margin drop are compared across all demonstration tokens, label tokens, and input tokens.

	(a) Recovery (%)			(b) Margin drop vs. control		
Model	All	Labels	Inputs	All	Labels	Inputs
SST-2						
Qwen2.5-7B	99.4	98.4	4.4	2.27	2.09	0.29
Mistral-7B	99.8	98.7	2.4	1.24	1.37	0.38
Llama-3.1-8B	96.2	95.1	3.2	0.63	0.76	0.23
TREC						
Qwen2.5-7B	99.4	98.8	3.7	1.04	0.89	0.47
Mistral-7B	98.5	91.1	11.1	1.37	1.35	0.14
Llama-3.1-8B	97.0	90.9	1.6	0.69	0.49	0.10

Note. Recovery is evaluated on examples with $|\Phi_r| \geq 0.5$. Matched positions outside the demonstration blocks form the control. Drop metrics are computed as control minus targeted attenuation, so larger values indicate stronger dependence on the selected positions.

Table 9(a) shows that the label positions recover nearly the full effect of the demonstration blocks. The input positions recover only a small fraction of this effect. Table 9(b) shows the same organization under targeted attenuation. The margin drop at the label positions is close to that of all demonstration positions, while the input positions show a much smaller change. Together with the pairing result in Table 8(a), these findings separate rule formation from its direct realization at the query decision. The rule is induced from relations across demonstrations, and its direct contribution is concentrated at the label positions.

D.5. Evidence from Multiple Sources

Our main evidence experiments place the relevant fact in a single support span, allowing us to isolate its effect on a factual choice. We next examine evidence use across multiple sources. One setting requires the model to combine information from several passages, while the other presents conflicting facts from two sources. In each setting, we identify which evidence positions affect the selected answer and test whether weakening them changes the factual choice.

Distributed evidence. We first study questions from MuSiQue (Trivedi et al., 2022) that require two reasoning steps. Each answer depends on a bridge passage and a passage that contains the final

fact. The paired prompts preserve the question, passage order, and all other text, while changing only the answer entity. We use the union of the two passages as the evidence positions and matched distractor positions as the control. This construction lets us compare the evidence needed to establish the answer with the evidence that acts at the final choice.

Table 10: **Evidence contributions under distributed multi-hop grounding. (a)** Grounded preference and pathway recovery. **(b)** Dependence on each supporting passage and its direct effect at the answer decision.

(a) Grounded preference and pathway recovery

	Grounded preference (%)			Matched effect	Pathway recovery
Model	Original	Counterfactual	No evidence	Median M_e	Recovery difference (nats)
Qwen2.5-7B	82.7	46.9	63.3	36.47	5.08 [4.0, 6.2]
Mistral-7B	89.8	43.9	69.4	35.11	5.07 [4.2, 6.0]
Llama-3.1-8B	98.0	50.0	76.5	26.08	4.30 [3.6, 5.0]

(b) Support dependence and decision-level effects

	Prompt dependence: margin drop (nats)		Decision effect: margin drop vs. control			Token flip (%)	
Model	Bridge removed	Answer removed	All evidence	Answer span	Bridge span	Targeted	Control
Qwen2.5-7B	1.98 [0.9, 3.1]	5.82 [4.5, 7.2]	5.89 [4.7, 7.1]	4.68 [3.7, 5.7]	0.09 [−0.4, 0.6]	11.2	1.0
Mistral-7B	1.27 [0.7, 1.9]	5.46 [4.4, 6.6]	3.98 [3.2, 4.8]	3.48 [2.8, 4.2]	0.14 [−0.1, 0.4]	8.2	0.0
Llama-3.1-8B	1.07 [0.6, 1.5]	4.50 [3.7, 5.3]	4.37 [3.6, 5.1]	3.42 [2.8, 4.0]	0.43 [0.1, 0.8]	12.2	0.0

Note. Recovery difference (nats) is the score recovery at the answer decision from evidence substitution relative to the matched distractor substitution. Larger margin drop indicates stronger dependence.

As shown in Table 10(a), the matched effect favors the supported answer for all three models, and substitution through the evidence positions recovers most of this effect relative to the control. Removing either passage also reduces the answer margin, showing that both passages are needed to establish the answer (Table 10(b)). Their direct effects at the final choice are not equal. Weakening the passage that contains the final fact causes a large margin drop in every model, whereas weakening the bridge passage has little direct effect in Qwen and Mistral and a much smaller effect in Llama. Thus, the answer depends on evidence distributed across the two passages, but the final choice depends mainly on the passage that supplies the final fact.

Competing evidence. We next consider contexts in which two sources state different entities for the same fact. The prompt marks one source as primary. The paired prompts exchange the entities stated by the two sources while preserving the question, passage order, and remaining text. A closed-book condition measures the preference that remains without either claim. We then weaken the two sources separately to test how each one affects the same factual choice.

Table 11(a) shows that context-conditioned preference varies across settings. The original and closed-book rates are nearly identical for Qwen, separate more for Mistral, and differ most clearly for Llama. The matched effect remains positive in all three models. Table 11(b) reveals a more consistent source-specific effect. Weakening the primary source lowers the margin of its answer, whereas weakening the competing source can raise it. The primary intervention also changes the selected token far more often than the control. Thus, the models distinguish the effects of the two sources, even though the designated source does not exert the same control over the final answer.

Taken together, the two settings show that evidence dependence can be distributed across sources, while direct influence and final control remain source-specific. The context determines which information is required, that information shapes the factual choice, and the resulting influence may or may not determine the final answer. These findings extend the evidence contribution account beyond a single support span while preserving the distinction between contextual influence and final behavior.

Table 11: **Evidence contributions under competing factual sources.** **(a)** Context-conditioned preference and pathway recovery. **(b)** Source-specific effects under targeted attenuation.

(a) Context-conditioned preference and pathway recovery

	Primary preference (%)			Matched effect		Pathway recovery
Model	Original	Swapped	Closed book	Median M_e	Positive M_e (%)	Recovery difference (nats)
Qwen2.5-7B	28.4	25.3	26.3	1.06	53.7 [43.2, 64.2]	4.99 [4.0, 6.0]
Mistral-7B	31.6	13.7	22.1	1.90	65.3 [55.8, 74.7]	4.63 [3.9, 5.3]
Llama-3.1-8B	34.7	5.3	20.0	4.03	81.1 [72.6, 88.4]	4.70 [4.1, 5.3]

(b) Source-specific causal effects

	Source-specific decision effect		Token flip (%)	
Model	Primary drop	Secondary gain	Primary	Control
Qwen2.5-7B	6.43 [5.2, 7.7]	5.30 [4.2, 6.5]	23.2	3.2
Mistral-7B	5.12 [4.3, 5.9]	3.91 [3.2, 4.6]	26.3	1.1
Llama-3.1-8B	6.30 [5.5, 7.1]	3.34 [2.8, 3.9]	31.6	0.0

Note. Primary drop is the margin decrease of the primary asserted candidate under targeted attenuation, relative to the matched control; Secondary gain is the margin increase after attenuating the competing source.

Table 12: **Role span behavior at semantic decisions.** **(a)** Role span influence at semantic decisions: the median within-response percentile of $|\hat{\delta}_t|$ at the decision position and at matched positions in other responses. **(b)** Position of the first generation change relative to the semantic decision.

	(a) Role span influence (%)		**(b) Position of generation changes**		
Model	Decision percentile	Matched control	Median offset	Before dec. (%)	At dec. (%)
Rule					
Qwen2.5-7B	**85.0**	67.4	$+0$	46.9	53.1
Mistral-7B	**85.7**	63.6	-1	55.9	44.1
Llama-3.1-8B	**90.0**	63.0	-5	55.3	44.7
Evidence					
Qwen2.5-7B	**92.5**	50.2	-8	86.0	8.0
Mistral-7B	**77.5**	51.2	-6	87.2	4.3
Llama-3.1-8B	**95.0**	50.5	-6	83.7	9.3

Note. **(a)** Both columns show the median within-response percentile of $|\hat{\delta}_t|$; Matched reads the same positions in other responses. **(b)** Offset is the first-divergence position minus the semantic-decision position (in tokens); negative values indicate an earlier change.

D.6. Role Span Influence around Key Decisions

So far, our analysis shows that role span contributions act through the output decisions required by each role. The same pattern extends to rules inferred from demonstrations and evidence drawn from multiple sources in Appendices D.4 and D.5. For rules, these decisions govern structure and selection; for evidence, they determine factual content in the response. One may wonder how the role effect is organized around these semantic decisions.

We answer this by comparing role span influence at the semantic decisions with matched output positions. As shown in Table 12(a), role span influence is consistently higher at the semantic decisions than at the matched positions. The same concentration appears when the decisions are specified from reference responses. These results show that semantic decisions form distinct points of role span influence in the output.

Step further, we ask how the generation change is positioned around these semantic decisions. We measure the first divergence relative to each decision position. As shown in Table 12(b), rule changes occur at or shortly before the decision itself, with 44%–53% of first divergences occurring exactly at that position. In contrast, evidence shows a different pattern: 84%–87% of first divergences occur before the factual value is emitted, with median offsets of several tokens. These results indicate that rule and evidence exhibit different positional patterns around their semantic decisions.

In short, the semantic decisions provide distinct points of role span influence, while generation changes follow different positional patterns around them. Functional roles therefore organize both the output decisions that carry role span influence and the local pattern of generation changes around those decisions.

D.7. Predicting Task Recovery

Appendix B.4 predicts the recovery of each naturally failing decision under role contribution amplification. We first compare these predictions with the subsequent intervention outcomes. Across the two roles, the predicted and observed decision outcomes agree in 79% of the cases, while the predicted amplification closely tracks the first observed crossing among recovered decisions.

Here, we show that the decision-level predictions can extend to task behavior. Following the role-specific decision organization identified in Section 5.2, we map the predicted decision outcomes to task recovery. For evidence, recovery of the factual choice determines the grounding prediction. For rules, the decision-level prediction is not directly comparable, since successful rule following requires sustained support across multiple structural and selection choices.

As shown in Table 13, the observed task recovery follows the predicted trend for evidence, while rule recovery stays near zero, as predicted. For evidence, task success increases from 25.2% to 58.3%, compared with a predicted increase from 23.3% to 69.9%. For rules, both remain near zero over the same range. Thus, we show that the first-order decision formulation can be used to quantitatively predict task recovery in both ICL and OOC hallucination.

Table 13: **Task-level recovery prediction under role contribution amplification.** Predicted versus observed task success across amplification strengths.

	Rule span		Evidence span	
β	Predicted (%)	Observed (%)	Predicted (%)	Observed (%)
0.5	0.0	0.0	23.3	25.2
1	0.0	1.3	26.2	32.0
2	0.0	2.5	38.8	47.6
4	0.0	6.3	52.4	53.4
8	0.0	3.8	64.1	55.3
16	6.3	6.3	69.9	58.3