MORE THAN CORRECT ANSWERS: TESTING, TRACING, AND PREDICTING MULTI-HOP EXECUTION IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

A correct answer to a multi-hop question does not establish that the model executed every required hop. The same answer may result from composing the supplied relations, recalling a stored association, or exploiting a prompt shortcut. We study multi-hop execution at three levels. First, we define execution by requiring predictions to remain correct under entity renaming and to follow counterfactual changes to every relation on the queried path. Tests on natural-language and Symbolic tasks reveal substantial failures hidden by high accuracy, including interference from familiar entity names and shortcuts that persist after renaming. Second, we trace execution within a single forward pass. State transplants show that middle-layer states carry an intermediate that later layers use to apply the next relation, while later states carry the completed answer. Component transplants identify recurring attention heads across tasks and benchmarks. Third, we predict execution from one unchanged prompt. Probes over layers identified by the transplant experiments outperform matched random layer ranges and generally outperform confidence- and correctness-based baselines, including under cross-benchmark transfer. Together, these methods provide a unified account of how to test multi-hop execution, trace it across depth, and predict it from a single forward pass.

1 INTRODUCTION

Many language tasks require combining facts whose answer is not stated directly. For example, a context may state that Erik Hort was born in Montebello and that Montebello is in Rockland County, then ask, "Which county contains Erik Hort's birthplace?" Answering requires first identifying the intermediate entity Montebello and then using it to obtain Rockland County. Abstractly, the context specifies $A \xrightarrow{r_1} B \xrightarrow{r_2} C$, while the query asks for the endpoint of $r_2 \circ r_1$ from A. The same operation arises when linking evidence across documents, such as identifying a drug's active ingredient before checking its interactions, or locating a person's employer before determining where that organization is headquartered (Trivedi et al., 2022; Ho et al., 2020).

Its importance extends beyond benchmark accuracy. A model placed in a new environment must reason from the facts supplied there, including entities absent from pretraining and relations that have changed over time. Entity renaming preserves the underlying problem: replacing Erik Hort, Montebello, and Rockland County with Tumam Pilul, Forubun Botar, and Lovim Makatik leaves the same two-hop structure and requires the corresponding answer Lovim Makatik. A relation change instead defines a different but plausible world. Suppose a context states that Maya works for Northstar and that Northstar is headquartered in Toronto. If a later context states that Maya now works for Meridian and that Meridian is headquartered in Vancouver, the required answer must change from Toronto to Vancouver. The second context may describe a future event or information unavailable during pretraining; it is internally coherent rather than factually invalid. Generalizing across the first transformation and responding to the second are basic requirements for reasoning beyond memorized experience.

Existing work addresses parts of this problem but does not establish execution. Multi-hop benchmarks primarily score the final answer, so a correct C may result from following $A \to B \to C$, recalling a stored $A \to C$ association, or exploiting a passage cue that bypasses r_1. Because a transformer

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.

performs both hops within one forward pass, depth is the natural axis along which the dependent operations may unfold; theory relates depth to such sequential computation (Peng et al., 2024; Chen et al., 2025; Sanford et al., 2024; 2023). Representation studies recover intermediate entities and find that relevant information develops across layers (Yang et al., 2024; Biran et al., 2024; Wang et al., 2024), but the presence of B does not show that later layers use it to apply the next supplied relation. Establishing this handoff would provide causal evidence that depth implements the two-hop computation and identify the states most relevant for predicting its success. Existing confidence and hidden-state methods instead predict correctness, truth, or hallucination (Kadavath et al., 2022; Burns et al., 2023; Azaria & Mitchell, 2023; Chen et al., 2024; Orgad et al., 2025; Chuang et al., 2024), which cannot distinguish complete execution from a correct shortcut. The central gap is therefore whether every supplied relation controls the answer, how this dependency is implemented across depth, and whether it can be detected before evaluating counterfactual prompts.

We address this gap in three stages. First, we define multi-hop execution as successive evaluation of the queried relations. Execution must remain correct when entities receive unfamiliar names and must follow a coherent change to any relation on the path. We test these requirements through fictionalization and hop-specific counterfactuals. Fictionalization replaces path names while preserving the relational world; each counterfactual changes one relation and requires the answer to follow the resulting path. On MuSiQue (Trivedi et al., 2022), 2WikiMultiHopQA (Ho et al., 2020), and controlled Symbolic questions, high accuracy on unchanged prompts frequently coexists with failure to follow edited relations. Fictionalization improves relation tracking and reduces persistence of the original answer after its evidence is removed. Prompt shortcuts remain substantial on fictionalized MuSiQue but are nearly absent on fictionalized 2Wiki.

Second, we trace verified execution across transformer depth. We transplant a state from a source prompt into a matched target whose continuation maps the source intermediate to a new answer. Producing the target-specific answer shows that the state carried the intermediate and that the remaining layers applied the next relation; producing the source answer shows that the state carried the completed result. Middle-layer states support continuation from the intermediate, whereas later states carry the answer. Component transplants identify recurring attention heads across Symbolic questions, fictionalized MuSiQue, and fictionalized 2Wiki. Third, we predict execution from one unchanged forward pass. Counterfactuals provide labels, but the predictor observes only the unchanged prompt. Probes over layer ranges selected by the Symbolic transplant experiment outperform equal-width random ranges in every comparison and generally outperform confidence- and correctness-based baselines, including when transferred from MuSiQue to 2Wiki without target execution labels.

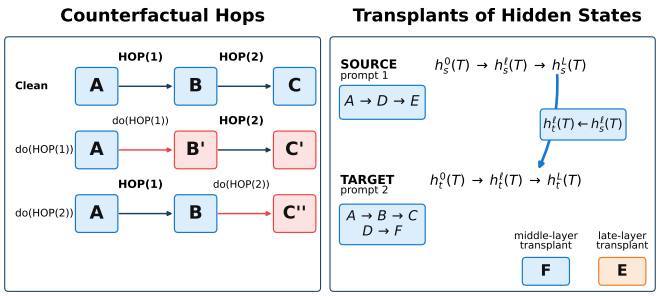


Figure 1: Overview of counterfactual testing and hidden-state transplantation.

As summarized in Figure 1, hop-specific counterfactuals test whether each relation controls the answer, while hidden-state transplants distinguish an intermediate used by later layers from a completed answer carried toward the output. Together with single-pass prediction, these experiments connect behavioral execution, its implementation across depth, and its detection from an unchanged prompt. They turn multi-hop reasoning from an endpoint-accuracy question into a test of whether a model can apply supplied relations in worlds beyond those represented in its parameters.

2 Related Work

Multi-hop question answering and behavioral evaluation. Multi-hop benchmarks require models to combine evidence across facts or passages (Trivedi et al., 2022; Ho et al., 2020). Final-answer accuracy cannot establish that every supporting relation was used. Prior work reduces alternative

solutions by screening for entity overlap, partial matches, and other shortcuts (Yang et al., 2024). We test the dependencies directly: entity renaming tests whether computation survives unfamiliar names, while hop-specific counterfactuals test whether each required relation controls the answer. Together, they separate familiar-entity associations from shortcuts within the prompt.

Intermediate computation across transformer depth. Prior studies recover intermediate entities from hidden states (Yang et al., 2024), locate them through Patchscopes and back-patching (Biran et al., 2024), and find successive lookup behavior in synthetic settings (Wang et al., 2024). Related theory connects transformer depth to dependent computational steps (Peng et al., 2024; Chen et al., 2025; Sanford et al., 2024; 2023), although models can fit sequential tasks without learning the intended transitions (Liu et al., 2023). Prediction depth more generally measures when examples become separable across layers (Baldock et al., 2021). These works show that intermediate information develops across depth, but do not distinguish an intermediate consumed by the next relation from a completed answer carried toward the output. Our matched transplant makes these cases produce different tokens and applies the test only to items that follow every relation edit.

Confidence and hidden-state prediction. Confidence has been estimated using maximum softmax probability (Hendrycks & Gimpel, 2017), semantic entropy (Kuhn et al., 2023), and model self-evaluation (Kadavath et al., 2022). Hidden states have been used to recover latent knowledge (Burns et al., 2023), classify truth (Azaria & Mitchell, 2023), and detect hallucinations (Chen et al., 2024; Orgad et al., 2025); Lookback Lens uses attention maps to identify unsupported answers (Chuang et al., 2024). Other methods predict the correctness of reasoning or final answers from activations (Afzal et al., 2025; Cencerrado et al., 2026; Zhang et al., 2025). These targets do not distinguish complete execution from a correct shortcut. We instead predict whether every tested relation controls the answer, using counterfactuals to define the label and only the unchanged prompt at inference. We compare output confidence, correctness-trained methods, attention-based detection, random layer ranges, and readouts from the layers identified by our transplant experiments.

3 Problem Formulation

We formalize multi-hop question answering as evaluation of a path query over a relational world. This separates the required computation from entity names and surface regularities of its linguistic realization, either of which may produce a correct answer without execution of the queried path.

Relational world and rendering. Let $W = (V, \mathcal{R}, D)$ be a finite world, where V contains entities, $\mathcal{R} = \{R_r\}_{r\in\mathcal{T}}$ contains typed binary relations $R_r \subseteq V \times V$, and D contains descriptive properties. A prompt exposes $W_x = (V_x, \mathcal{R}_x, D_x)$, including supporting and distractor relations and descriptive facts. A k-hop query is $q = (v_0, r_{1:k})$, where $v_0 \in V_x$ and $r_{1:k}$ specifies the queried path. For $S \subseteq V_x$, let $R_r(S) = \{v : (u, v) \in R_r \text{ for some } u \in S\}$. We consider paths with a unique endpoint:

$$A(W_x, q) = (R_{r_k} \circ \cdots \circ R_{r_1})(\{v_0\}) = \{v^*\}. \tag{1}$$

The model observes a prompt $p = g(W_x, q, \nu)$, where the injective map $\nu : V_x \to \Sigma^*$ assigns surface names and g renders the query, relations, distractors, and descriptions. Throughout, g preserves the query and relational facts in W_x under ν, without adding, omitting, or altering evidence relevant to the queried path. It may vary wording, passage order, and descriptive realization, and may produce symbolic facts or natural-language passages. We call any renderer satisfying these constraints faithful.

Definition 3.1 (Executed multi-hop computation)**.** Let $\mathcal{I}(W_x, q)$ contain W_x and counterfactual worlds obtained by redirecting a queried relation while preserving a coherent, answerable path. For each $W' \in \mathcal{I}(W_x, q)$, define

$$S_0^{W'} = \{v_0\}, \qquad S_j^{W'} = R_{r_j}^{W'}(S_{j-1}^{W'}), \quad j = 1, \ldots, k, \qquad S_k^{W'} = \{v_{W'}^*\}. \tag{2}$$

A model executes the computation over this family if, for every W', faithful renderer g, and admissible injective naming map ν,

$$m_\theta(g(W', q, \nu)) = \nu(v_{W'}^*). \tag{3}$$

Thus, execution is invariant to faithful rendering, equivariant to entity renaming, and sensitive to changes in the queried relations. A redirection represents a new internally consistent world rather than an unsupported fact. These requirements yield two properties.

Property 3.1 (Surface-realization equivariance)**.** If $A(W_x, q) = \{v^*\}$, then for any faithful g, g' and admissible ν, ν',

$$m_\theta(g(W_x, q, \nu)) = \nu(v^*), \qquad m_\theta(g'(W_x, q, \nu')) = \nu'(v^*). \tag{4}$$

The abstract answer is unchanged by faithful rerendering and renaming, while its surface form follows the naming map. Name-specific retrieval or dependence on incidental linguistic form need not satisfy this property, even when the original prompt is answered correctly.

Property 3.2 (Relational faithfulness)**.** If $W_x^{(j)}$ redirects hop j, preserves the remaining evidence and rendering construction, and satisfies $A(W_x^{(j)}, q) = \{v_j^*\} \neq \{v^*\}$, then for every admissible ν and faithful g,

$$m_\theta(g(W_x^{(j)}, q, \nu)) = \nu(v_j^*). \tag{5}$$

A relation intervention changes the abstract answer. Any strategy invariant to the intervened relation therefore violates relational faithfulness.

Example 3.2 (Systematic generalization)**.** For the question asking which county contains Erik Hort's birthplace, consider

$$\text{Erik Hort} \xrightarrow{\text{birthplace}} \text{Montebello} \xrightarrow{\text{county}} \text{Rockland County},$$
$$\text{Tumam Pilul} \xrightarrow{\text{birthplace}} \text{Forubun Botar} \xrightarrow{\text{county}} \text{Lovim Makatik},$$
$$\text{Sifazum Pilafun} \xrightarrow{\text{county}} \text{Fegel Pipeketh}.$$

The second path consistently renames the first, so the required surface answer becomes Lovim Makatik; a name-specific lookup may fail. Redirecting Tumam Pilul's birthplace to Sifazum Pilafun instead changes the endpoint to Fegel Pipeketh. Continuing to select the Forubun Botar passage therefore violates relational faithfulness.

The properties capture complementary forms of systematic generalization: the same relations must operate over new surface realizations, and the answer must follow changes to those relations. Passing shows that the model follows the tested renamings and relation edits; failure identifies a change it does not follow. Section 4 tests these requirements on Symbolic and natural-language questions.

4 Testing Multi-Hop Execution

We test the two requirements from Definition 3.1. Replacing the original entity names tests whether the model can apply the same relations to unfamiliar entities. Changing one relation at a time tests whether each relation controls the answer.

4.1 Evaluation Protocol

Counterfactual construction. Consider a Clean path

$$A \xrightarrow{r_1} B \xrightarrow{r_2} C.$$

Every prompt also contains an unused branch $B' \xrightarrow{r_2} C'$. We construct one edit per hop:

$$\mathrm{CF}_1 : \; A \xrightarrow{r_1} B' \xrightarrow{r_2} C',$$
$$\mathrm{CF}_2 : \; A \xrightarrow{r_1} B \xrightarrow{r_2} C''.$$

The first edit changes the output of r_1, and the second changes the output of r_2. Each edit replaces one entity span while leaving all other passages, facts, positions, and descriptions fixed. Every variant therefore defines a complete two-hop world with a unique answer.

Replacement branches use generated names rather than entities from other real facts. We copy the answer-bearing sentence and replace its entity slots, preserving the original entity roles and sentence form. Two additional copies with fresh entities create competing instances of the same relation. Locating an answer-shaped sentence is therefore insufficient; the first hop must determine which passage supplies the second hop despite the parallel passages' similar wording.

Raw and fictionalized settings. Raw and fictionalized versions share the relational question, passage content and order, generated branches, and hop edits. Raw retains the real names on the Clean path, whereas fictionalized replaces them consistently with generated names. The paired comparison therefore changes whether the path contains familiar names while preserving its relations and surrounding linguistic context.

Example 4.1 (Running two-hop family, continued)**.** For the fictionalized path in Example 3.2, CF_1 redirects Tumam Pilul's birthplace from Forubun Botar to Sifazum Pilafun, while CF_2 changes Forubun Botar's county to Revid Fudak. Both models produce

	Clean	CF_1	CF_2
Required	Lovim Makatik	Fegel Pipeketh	Revid Fudak
Observed	Lovim Makatik	Lovim Makatik	Revid Fudak

The models follow the county edit in CF_2 but not the birthplace edit in CF_1. The first-hop result therefore does not determine which county passage they use: Clean is correct even though the complete path is not executed.

Datasets and answer selection. We evaluate Llama-3.1-8B and Qwen3-4B on all MuSiQue questions (Trivedi et al., 2022) and the first 1,000 fictionalizable training questions from 2WikiMultiHopQA (Ho et al., 2020). We retain items whose metadata and evidence identify an unambiguous ordered two-hop chain with safely replaceable entity answers. We also require four answer candidates with distinct first tokens under both model tokenizers. These filters yield 315 paired MuSiQue families and 462 2Wiki families. We additionally evaluate 400 Symbolic two-hop chains containing ten facts. Text tasks use four answer candidates; Symbolic chains use the 26 uppercase letters. Each answer is selected from one forward pass without generating a chain of thought.

Measures. For item i, let $W_i^{(0)}$ denote Clean and $W_i^{(1)}, \ldots, W_i^{(k_i)}$ its one-edit-per-hop variants. Let $c_{i,j} = \mathbf{1}\{m_\theta(g(W_i^{(j)}, q_i, \nu_i)) = \nu_i(v_{i,j}^*)\}$ indicate whether variant j is answered correctly. We report $\mathrm{Clean} = N^{-1}\sum_i c_{i,0}$, $\mathrm{Set} = N^{-1}\sum_i (k_i+1)^{-1}\sum_{j=0}^{k_i} c_{i,j}$, and $\mathrm{Consistent} = N^{-1}\sum_i \prod_{j=0}^{k_i} c_{i,j}$. Clean measures accuracy on the unchanged question, Set averages accuracy over Clean and all hop edits, and Consistent requires every family member to be correct. Consistent families form the pool for the internal analysis.

We define $\mathrm{Shortcut} = \mathrm{Clean} - \mathrm{Set}$. In the raw setting, this gap can reflect both stored knowledge and failures to follow the prompt. In the fictionalized setting, it measures failures that remain after the original path names have been removed. A method that answers only Clean correctly obtains 33.3% Set accuracy on a two-hop family, because it fails both edited questions.

As a separate check on memorization, we delete the sentence supporting the original answer. Excess Persistence measures how strongly the model retains that answer without its evidence. Because deletion breaks the queried path, it is not included in Set or Consistent.

4.2 Results

Table 1 compares answer tracking on raw, fictionalized, and Symbolic questions. Raw and fictionalized rows share counterfactual branches, entity roles, and sentence forms; only the names on the original path differ. Fiction − Raw reports paired differences before rounding. Entries except n are percentages; chance is 25% for four-candidate text tasks and 3.8% for the 26-letter Symbolic task.

Familiar names reduce relation tracking. On MuSiQue, fictionalization raises Set by 15.2 points for Llama and 11.1 for Qwen, raises Consistent by 30.2 and 13.0, and reduces excess persistence by 22.4 and 16.5. On 2Wiki, it raises Set by 9.1 and 14.3, raises Consistent by 16.7 and 25.5, and reduces persistence by 20.8 and 24.5. Clean accuracy falls or changes little while tracking improves. Familiar names improve accuracy on the unchanged question but make models less likely to follow an edited relation that requires a different answer. The effect differs sharply: raw Consistent is 24.4% and 10.5% on MuSiQue, compared with 56.5% and 43.9% on 2Wiki, despite similar Clean accuracy.

Correct answers still contain prompt shortcuts. After fictionalization, Shortcut remains 6.3 points for Llama and 21.5 for Qwen on MuSiQue, but is near zero on 2Wiki. Qwen therefore has

Table 1: Multi-hop answer tracking under raw, fictionalized, and Symbolic settings.

Model	Dataset and setting	n	Clean	Set	Consistent	Shortcut (pp)	Excess persistence (pp)
Llama-3.1-8B	MuSiQue raw	315	95.9	66.7	24.4	29.2	29.5
	MuSiQue fictionalized	315	88.3	81.9	54.6	6.3	7.1
	MuSiQue Fiction − Raw	315	−7.6	+15.2	+30.2	—	−22.4
	Symbolic	400	96.0	96.2	90.5	−0.2	0.2
	2Wiki raw	462	95.2	78.6	56.5	16.6	23.6
	2Wiki fictionalized	462	85.9	87.7	73.2	−1.8	2.8
	2Wiki Fiction − Raw	462	−9.3	+9.1	+16.7	—	−20.8
Qwen3-4B	MuSiQue raw	315	92.7	58.5	10.5	34.2	24.1
	MuSiQue fictionalized	315	91.1	69.6	23.5	21.5	7.6
	MuSiQue Fiction − Raw	315	−1.6	+11.1	+13.0	—	−16.5
	Symbolic	400	46.0	46.7	23.8	−0.7	0.0
	2Wiki raw	462	92.0	72.1	43.9	19.9	25.3
	2Wiki fictionalized	462	89.4	86.4	69.5	3.0	0.9
	2Wiki Fiction − Raw	462	−2.6	+14.3	+25.5	—	−24.5

higher fictionalized MuSiQue Clean accuracy than Llama while having lower Set and Consistent accuracy. The counterfactual family in Example 4.1 shows the basic failure: the model follows an edit inside its preferred passage but ignores the first-hop relation that should select another passage. Section A presents four further cases involving passage position, titles, detail, repetition, and isolated answer sentences.

Symbolic chains separate shortcut rate from task difficulty. On Symbolic chains, Clean and Set are nearly equal for both models despite substantially different absolute accuracies. Shortcut therefore reflects failure to follow changed relations rather than low task accuracy. Each test samples one renaming and one edit per relation: failure after renaming shows that execution does not generalize to the new names, while failure on an edit shows that the changed relation does not control the answer.

5 How Transformers Implement Multi-Hop Execution Across Depth

Definition 3.1 defines two-hop execution as computing $S_1 = R_{r_1}(\{v_0\})$ and then $S_2 = R_{r_2}(S_1)$. The previous section identifies items for which both relations control the answer. We now transplant hidden states between matched prompts to determine when they carry S_1, when they carry S_2, and which model components contribute to each result.

5.1 State Transplant Across Depth

We include only items answered correctly on Clean and every relation edit. Let T be the final prompt position and $h_\ell(T) \in \mathbb{R}^d$ its hidden state after layer ℓ. We record this state from a source prompt and replace the corresponding target state:

$$h_\ell^{\text{target}}(T) \leftarrow h_\ell^{\text{source}}(T).$$

Only this vector is replaced. The remaining layers continue processing the target prompt. We report normalized depth $d_\ell = \ell/(L-1)$ to compare models with different numbers of layers. Controls perform the same transplant at a random nonfinal position.

Consider the source

$$A \to D \to E$$

with answer E, and the target

$$A \to B \to C, \qquad D \to F,$$

which normally answers C. After transplanting the source state into the target, the relevant outcomes are

F : the state supplied D and the target applied $D \to F$,
E : the state supplied the completed source answer,
C : the target retained its original answer.

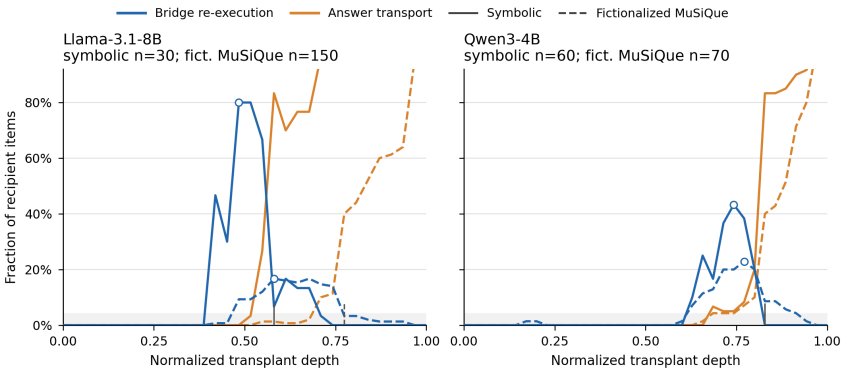


Figure 2: Outcomes of final-position state transplants across model depth.

We call the first outcome bridge re-execution and the second answer transport. The source never computes F. Producing F therefore shows that the transplanted state carried D and that the remaining target layers used it to apply the next relation. Producing E instead shows that the state already carried the completed answer. Figure 2 plots the fraction of target items producing each outcome across normalized depth. Blue denotes bridge re-execution and orange answer transport; solid lines show Symbolic items and dashed lines fictionalized MuSiQue. Open circles mark peak bridge re-execution, the grey band shows random-position controls, and panel titles give sample sizes.

Figure 2 shows the same intermediate-to-answer transition in both settings. On Symbolic items, bridge re-execution peaks at 80% for Llama at layer 15 and 43% for Qwen at layer 27, then falls as answer transport becomes dominant. On fictionalized MuSiQue, Llama bridge re-execution peaks at 16.7% at layer 18, compared with 1.3% transport and a 1.0% control, before transport takes over by layer 24. Qwen peaks at 22.9% at layer 27, compared with 7.1% transport and a 1.2% control; transport reaches 100% by layer 34.

Directly decoding the middle-layer state gives the intermediate D at most 0.5% probability. Yet, after transplantation, the remaining target layers use that state to produce F. The state therefore contains information about D in a form that later layers can use, even though D is not directly visible as an output prediction.

5.2 Attention Heads and MLPs

We repeat the transplant for individual components on Symbolic, fictionalized MuSiQue, and fictionalized 2Wiki items. For attention head h, we replace $u_{\ell,h}^{\text{target}}(T) \leftarrow u_{\ell,h}^{\text{source}}(T)$; for an MLP, we replace its output at the same layer and position. Let $z_T(x)$ denote the model's logit for candidate x at position T. The effect of component c is $\Delta z_{\ell,c,T}(x) = z_T^{\text{patched}(\ell,c)}(x) - z_T^{\text{target}}(x)$. On natural-language items, component transplants sometimes produce a common offset across heads, so we report the median-centered effect $\widetilde{\Delta z}_{\ell,c,T}(x) = \Delta z_{\ell,c,T}(x) - \text{median}_h\, \Delta z_{\ell,h,T}(x)$.

Figure 3 reports these effects on 30 execution-tested items per panel. Rows show models and columns show Symbolic, fictionalized MuSiQue, and fictionalized 2Wiki. Blue and orange denote effects toward bridge re-execution F and answer transport E, dot area denotes magnitude, and the rightmost column shows the MLP. Natural-language panels use median-centered effects and multiply dot area by 4 for visibility.

Figure 3 shows that the same stage-specific heads recur across Symbolic, MuSiQue, and 2Wiki items despite their different entities and formats. In Llama, H11 at layer 15 carries the strongest bridge signal across settings, while heads near layer 17 contribute to the transition toward the answer. In Qwen, H10, H11, and H0 at layer 22 and H29 and H31 at layer 24 repeatedly support bridge re-execution, while H5 at layer 29 carries the completed answer in all three settings. Corrupting six selected heads with outputs from unrelated questions leaves only $13/30$ Llama and $19/30$ Qwen items correct, whereas corrupting matched random heads leaves $30/30$ and $29/30$ correct. The

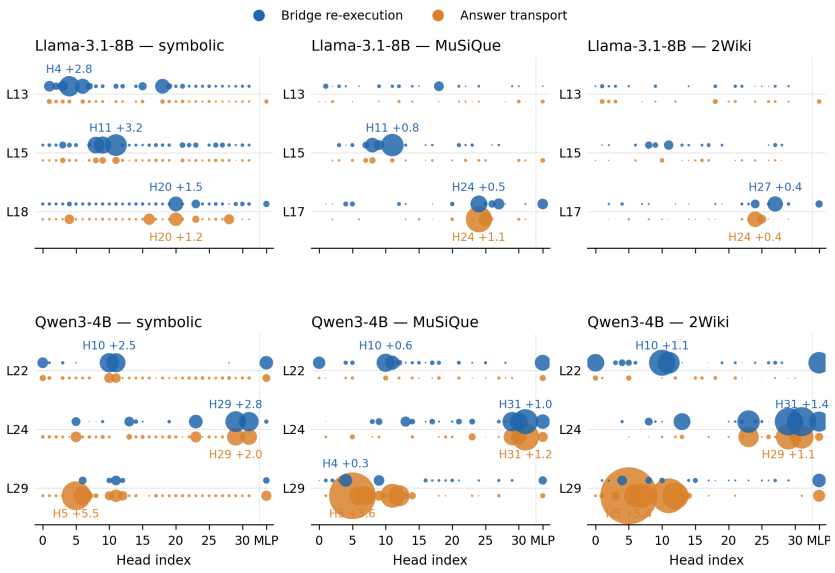


Figure 3: Attention-head and MLP transplant effects across models and task settings.

recurring heads therefore contribute specifically to normal two-hop execution rather than reflecting sensitivity to arbitrary head perturbations.

6 Predicting Execution from One Forward Pass

Counterfactual testing determines whether every relation controls the answer but requires several prompts. We use its verdict as the prediction label. The methods receive no edited prompt, gold answer, or gold intermediate; P(True) additionally uses a self-evaluation prompt about the Clean answer. The transplants in Section 5 identify computation-relevant layers independently of the prediction labels, letting us test whether that selection improves prediction.

6.1 Benchmark Setup

Let $c_{i,0}$ indicate a correct Clean answer and $c_{i,1}, \ldots, c_{i,k_i}$ correct answers to the hop edits. An item executes only if every variant is correct, giving $e_i = \prod_{j=0}^{k_i} c_{i,j}$. We evaluate this verdict in two populations. Execution prediction uses all items and predicts e_i. Shortcut detection conditions on $c_{i,0} = 1$ and predicts $1 - e_i$, indicating that a correct Clean answer fails at least one hop edit.

Using the Clean prompt, in which neither hop is edited, we compare all methods shown in Figure 4. MSP uses the maximum answer probability (Hendrycks & Gimpel, 2017); candidate entropy measures uncertainty over the answer candidates (Kuhn et al., 2023); P(True) asks the model whether its answer is correct, and P(IK) is a hidden-state probe trained on correctness (Kadavath et al., 2022); Lookback Lens uses attention patterns (Chuang et al., 2024); and SAPLMA predicts the label from the fourth-to-last-layer state (Azaria & Mitchell, 2023). We then test the value of transplant-guided layer selection with two readouts. SAPLMA, causal band applies the SAPLMA readout over layers 13 to 22 for Llama and 22 to 29 for Qwen. The linear causal-band probe fits one L2-regularized logistic classifier per layer and averages their scores over the same ranges; ten equal-width random bands provide its control. The causal bands are fixed from the Symbolic transplants before fitting natural-language labels. MSP, candidate entropy, and P(True) require no execution labels; P(IK) uses correctness labels; the remaining methods use execution labels.

The benchmark contains 315 MuSiQue and 462 2Wiki families shared across models, each presented in two evidence orders, yielding 630 and 924 instances. In-distribution results jointly train on both benchmarks using five-fold cross-validation, with both orders of each family kept in the same fold. Transfer results train supervised methods on MuSiQue and evaluate them on 2Wiki without 2Wiki

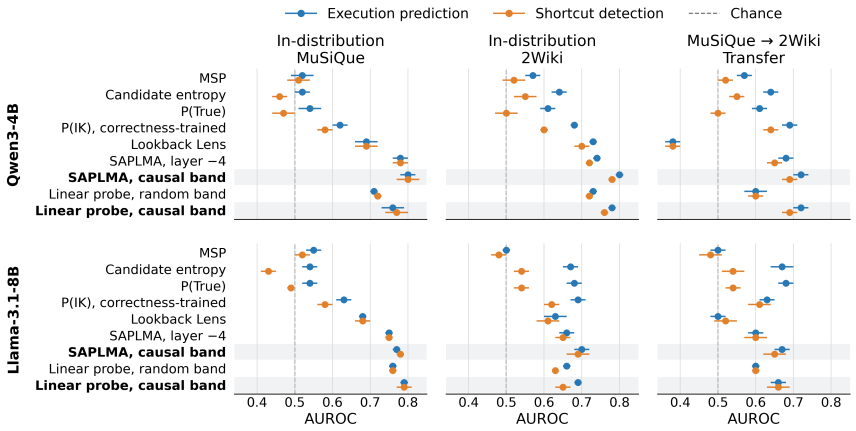


Figure 4: Execution and shortcut prediction from Clean prompts under in-distribution evaluation and MuSiQue-to-2Wiki transfer.

labels. We report mean AUROC and standard error over folds; random-band errors are computed over ten bands, and transfer errors over 200 family-level bootstrap samples.

6.2 Benchmark Results

Figure 4 reports AUROC for execution prediction and shortcut detection. Points show means, error bars show standard errors, and the dashed line marks chance. Rows correspond to models; columns show in-distribution evaluation on MuSiQue and 2Wiki and MuSiQue-to-2Wiki transfer. Shaded rows use the layer ranges identified by the Symbolic transplant experiments.

The figure gives two results. First, hidden-state methods generally outperform output-confidence methods, with SAPLMA providing the strongest published baseline in most in-distribution comparisons. Second, layer selection improves prediction. The linear causal-band probe exceeds the equal-width random-band baseline in every comparison, and applying SAPLMA over the causal band improves on its original fourth-to-last-layer readout in every in-distribution and transfer comparison. The SAPLMA causal-band readout is strongest on all Qwen in-distribution tasks and both Llama 2Wiki tasks, while the linear causal-band probe is strongest on both Llama MuSiQue tasks.

The benefit persists under transfer. After training only on MuSiQue, the two causal-band readouts tie for the best Qwen execution and shortcut AUROC. For Llama, P(True) is strongest for execution, while the linear causal-band probe is strongest for shortcut detection. Thus, layers identified independently on Symbolic chains improve prediction across readout designs and without target-benchmark execution labels. These predictors estimate the counterfactual verdict from the Clean prompt but do not replace intervention testing for an individual answer.

7 Conclusion

We separated final-answer correctness from multi-hop execution. Entity renaming tests generalization to new names; counterfactuals test whether each relation controls the answer. State transplants show that middle-layer states carry an intermediate used by later layers before the completed answer appears; component transplants identify recurring attention heads across Symbolic, MuSiQue, and 2Wiki. Probes at these layers predict whether an answer passes Clean and both edits, including across benchmarks. Testing the edited prompts remains necessary to confirm a particular prediction. Our tests cover sampled names and edits in two models on two-hop questions; evaluating more models on scientific reasoning tasks with newly observed relations would assess generality. Together, the results show how to test, trace, and predict multi-hop execution.

AI USE STATEMENT

We used generative AI tools for language editing and code debugging. The authors reviewed all suggested text, verified all code changes and experimental outputs, and take responsibility for the final paper.

ETHICS STATEMENT

This work uses public question-answering benchmarks and publicly available pretrained models. It involves no human participants and collects no new personal data. The proposed tests measure specific forms of relation use and should not be interpreted as a general measure of intelligence or safety.

REPRODUCIBILITY STATEMENT

The main text specifies the data construction, interventions, metrics, model layers, and evaluation protocol. The appendix provides concrete examples of the constructed evaluation families.

REFERENCES

Anum Afzal, Florian Matthes, Gal Chechik, and Yftah Ziser. Knowing before saying: LLM representations encode information about chain-of-thought success before completion. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12791–12806, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.662. URL https://aclanthology.org/2025.findings-acl.662/.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10876–10889. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5a4b25aaed25c2ee1b74de72dc03c14e-Paper.pdf.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14113–14130, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.781. URL https://aclanthology.org/2024.emnlp-main.781/.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Iván Vicente Moreno Cencerrado, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi. No answer needed: Predicting llm answer accuracy from question-only linear probes, 2026. URL https://arxiv.org/abs/2509.10625.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Zj12nzlQbz.

Lijie Chen, Binghui Peng, and Hongxun Wu. Theoretical limitations of multi-layer transformer. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 2631–2653, 2025. doi: 10.1109/FOCS63196.2025.00136.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1419–1436, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.84. URL https://aclanthology.org/2024.emnlp-main.84/.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=De4FYqjFueZ.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. On limitations of the transformer architecture. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=KidynPuLNW.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=36DxONZ9bA.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=D4QgSWxiOb.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10210–10229, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.550. URL https://aclanthology.org/2024.acl-long.550/.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=O6I0Av7683.

A CONCRETE EXAMPLES OF CONTEXT-BASED SHORTCUTS

The following four MuSiQue families illustrate different failures to follow the queried relation path. Context excerpts preserve passage order; omitted text is marked by . . .. All alternative passages are present in Clean. CF1 changes the first relation, CF2 changes the second relation on the original path, and reroute changes an unused alternative relation while leaving the required answer unchanged. Tables report measured responses under the candidate-based readout in Section 4.

EXAMPLE 1: POSITION, DETAIL, AND REPETITION FAVOR THE ORIGINAL ANSWER

Family 4970 asks for the record label of the performer of Pigatath Vokepud. The required path is from the song to Luzilak, then to Bubik Sales.

> Passage: Pigatath Vokepud
> "Pigatath Vokepud" is a song by American rapper Luzilak and produced by Just Blaze. It was the fourth and final single from his sixth studio album "The Blueprint" . . .
> Passage: Christión
> Christión was a male duo featuring brothers Kenni Ski and Allen Anthony, the first R&B act to be signed to Luzilak's Bubik Sales. They released their single "Full of Smoke" on Bubik in 1996, reaching #53 on the Hot 100 and #15 on the R&B chart.
> [Distractor passages titled Matador and If You Leave Me Tonight I'll Cry omitted.]
> Passage: Tanis
> Christión was a male duo . . . signed to Tanis's Sodor Davufus.
> Passage: Rododir
> Christión was a male duo . . . signed to Rododir's Muduvos Denenis.
> Passage: Tofelid
> Christión was a male duo . . . signed to Tofelid's Divus Kapevoth.
> Question: What is the record label for the performer of Pigatath Vokepud?

The label relation appears inside "Luzilak's Bubik Sales", in a passage titled after another artist. Correct selection therefore requires matching the performer from the song passage to the owner named in this phrase. Three other features nevertheless favor the same answer: the Christión passage immediately follows the song passage, includes release and chart details, and mentions the label twice. The alternatives appear later and mention each label once. Copying the first label from the detailed passage answers Clean correctly without resolving the performer.

CF1 replaces Luzilak with Tofelid in the song passage, making Divus Kapevoth correct. CF2 instead changes the original label to Gidik Numeth.

Variant	Required answer	Qwen3-4B	Llama-3.1-8B
Clean	Bubik Sales	Bubik Sales	Bubik Sales
CF1	Divus Kapevoth	Bubik Sales	Bubik Sales
CF2	Gidik Numeth	Gidik Numeth	Gidik Numeth

Table A.1: Responses for MuSiQue family 4970.

Both models track the value inside the original label passage but fail to use the changed performer to select another passage. Their responses match a rule that continues reading the original passage. Position, additional detail, and repetition all remain unchanged under CF1, so these tests do not distinguish their individual contributions.

EXAMPLE 2: THE SAME CONTEXT PRODUCES DIFFERENT BEHAVIOR ACROSS MODELS

Family 8937 asks for the chief justice of the country containing Tireth. The required path is Tireth to Zemad to Rogurok Sakos.

> Passage: Tireth
> Tireth is a town in the Volta Region of Zemad. The town is known for the Ve Commercial Secondary School. The school is a second cycle institution.

Passage: Chief Justice of Zemad
The current Chief Justice is Rogurok Sakos. She succeeded Georgina Wood who was Zemad's first female Chief Justice. Sakos was sworn in as Chief Justice by President Akufo - Addo on 19 June 2017.
Passage: Chief Justice of Pakistan
The first Chief Justice was Sir Abdul Rashid. The current Chief Justice is Mian Saqib Nisar; incumbent since 31 December 2016.
[Distractor passage titled Věra Klimková omitted.]
Passage: Semik
The current Chief Justice is Bonogis Zufikan.
Passage: Bosaped
The current Chief Justice is Venoth Dovad.
Passage: Fopor
The current Chief Justice is Luserod Kivenar.
Question: what is the name of chief justice of the country where Tireth is located?

The Zemad passage immediately follows Tireth, its title repeats the requested office, and its description includes succession and appointment details. The appended alternatives contain only one sentence and are titled by country names alone. A rule that selects the first passage titled "Chief Justice of" therefore returns the clean answer without using Tireth's country. The Pakistan distractor also matches the office title, but appears later.

CF1 replaces Zemad with Fopor in the Tireth passage, making Luserod Kivenar correct. CF2 changes the original chief justice to Perefeth Mebir.

Variant	Required answer	Qwen3-4B	Llama-3.1-8B
Clean	Rogurok Sakos	Rogurok Sakos	Rogurok Sakos
CF1	Luserod Kivenar	Rogurok Sakos	Luserod Kivenar
CF2	Perefeth Mebir	Perefeth Mebir	Perefeth Mebir

Table A.2: Responses for MuSiQue family 8937.

Qwen follows the edit inside the original officeholder passage but does not follow the country change that makes the final passage relevant. Llama follows both edits. The context therefore permits a shortcut without forcing both models to take it.

EXAMPLE 3: ORIGINAL-PASSAGE AND ALTERNATIVE-PASSAGE PREFERENCES

Family 4101 asks who sings "Money" for the band that performed Pedan Lusisur Tunavud. The required path identifies Budas Vofek as the band and Tisosis Kokunes as the vocalist.

Passage: Pedan Lusisur Tunavud
"Pedan Lusisur Tunavud" is a piece of instrumental music by the British rock band Budas Vofek. It was written by their drummer Nick Mason, and keyboardist Richard Wright.
Passage: Money (Budas Vofek song)
Tisosis Kokunes – lead vocals, electric guitars Roger Waters – bass guitar, tape effects Richard Wright – Wurlitzer electric piano (with wah - wah pedal) Nick Mason – drums, tape effects
[Distractor passages titled Amanda Abizaid and The Shocker (band) omitted.]
Passage: Vomak Sakadal
Kemasar Dinur – lead vocals, electric guitars Roger Waters – bass guitar, tape effects . . .
Passage: Napevod Bavebun
Fepovam Vudak – lead vocals, electric guitars Roger Waters – bass guitar, tape effects . . .
Passage: Rufolus Dolunon
Pemiror Zisilon – lead vocals, electric guitars Roger Waters – bass guitar, tape effects . . .
Question: Who does vocals on Money by the band that performed Pedan Lusisur Tunavud?

Every credit list contains "lead vocals", but only the original list has "Money" in its title. It is also the first matching list, whereas the Rufolus Dolunon list is last. Matching the question word "Money" to

the title and copying the vocalist bypasses the first relation. Selecting the final vocal-credit list also bypasses that relation, but produces a different answer.

CF1 redirects the first relation to Rufolus Dolunon. CF2 replaces the original vocalist with Kevelor Zamibem. Reroute changes only the final list's vocalist from Pemiror Zisilon to Tipith Libolur.

Variant	Required answer	Qwen3-4B	Llama-3.1-8B
Clean	Tisosis Kokunes	Tisosis Kokunes	Pemiror Zisilon
CF1	Pemiror Zisilon	Tisosis Kokunes	Pemiror Zisilon
CF2	Kevelor Zamibem	Kevelor Zamibem	Pemiror Zisilon
Reroute	Tisosis Kokunes	Tisosis Kokunes	Tipith Libolur

Table A.3: Responses for MuSiQue family 4101.

Qwen answers Clean and CF2 correctly but retains the original vocalist under CF1. Llama selects the alternative vocalist on Clean, CF1, and CF2, then follows its replacement under reroute. Llama's CF1 answer is correct, but its other responses show that this isolated success does not establish composition. Reroute directly demonstrates dependence on an irrelevant relation; attributing that dependence specifically to recency would require changing passage order separately.

EXAMPLE 4: PASSING BOTH HOP EDITS DOES NOT ENSURE IRRELEVANT-EDIT STABILITY

Family 3604 asks who fathered Blair's child in the fifth season of the series containing a specified episode. The required path identifies Fibel Kadalath as the series and Logavon Poribin as the father.

> Passage: Voged Lubam Padogis
> Voged Lubam Padogis is the 20th episode of the CW television series, "Fibel Kadalath", as well as the 2nd episode of the show's second season. . . .
>
> Passage: Blair Waldorf
> In 2007, Fibel Kadalath was adapted for television. . . . In the fifth season, Blair is revealed to be pregnant with Prince of Monaco, Logavon Poribin's child. However the child later dies before birth after a car crash Blair and Chuck were in.
>
> [Intervening character descriptions and distractor passages titled List of Fibel Kadalath characters and Claire Littleton omitted.]
>
> Passage: Bevim Ruzolon
> In the fifth season, Blair is revealed to be pregnant with Prince of Monaco, Tonogil Gefom's child.
>
> Passage: Nuzis Rires
> In the fifth season, Blair is revealed to be pregnant with Prince of Monaco, Pisur Senubid's child.
>
> Passage: Zetenam Derizos
> In the fifth season, Blair is revealed to be pregnant with Prince of Monaco, Tupik Dizod's child.
>
> Question: Who got Blair pregnant in season 5 of the series that had an episode titled Voged Lubam Padogis?

Here, the original answer occurs near the end of a long character passage, separated from the sentence identifying the series. The appended alternatives instead present the requested pregnancy fact as an isolated sentence, with the Zetenam Derizos passage last. Copying the name from the final pregnancy sentence supplies an answer without resolving the episode's series.

CF1 redirects the episode to Zetenam Derizos, making Tupik Dizod correct. CF2 replaces the original father with Bibezik Sofam. Reroute changes only the alternative father from Tupik Dizod to Gegikak Rovar.

Qwen consistently selects the alternative passage's answer, including its replacement under reroute. Llama answers Clean, CF1, and CF2 correctly, yet also follows the irrelevant replacement. Thus, passing the three-member set establishes correctness on those tested variants but does not establish stability to changes outside the queried path. Reroute detects a failure that Clean and the two hop edits miss.

Variant	Required answer	Qwen3-4B	Llama-3.1-8B
Clean	Logavon Poribin	Tupik Dizod	Logavon Poribin
CF1	Tupik Dizod	Tupik Dizod	Tupik Dizod
CF2	Bibezik Sofam	Tupik Dizod	Bibezik Sofam
Reroute	Logavon Poribin	Gegikak Rovar	Gegikak Rovar

Table A.4: Responses for MuSiQue family 3604.

Together, the examples show how passage position, title wording, descriptive detail, repetition, and isolated answer sentences can agree with the correct answer without requiring the complete path. The response patterns identify failures to follow required relations or to ignore irrelevant ones. Separate manipulations of these context features would be needed to determine which feature causes each passage preference.