KNOWLEDGE INTERACTION IN HYPERNETWORK-BASED CONTINUAL LEARNING: CONFLICT AND COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning aims to adapt pretrained large language models (LLMs) to evolving data while preserving their pretrained capabilities. A central challenge is catastrophic forgetting: adaptation to new data can impair performance on previously learned knowledge. Since continual learning commonly changes model parameters, the model's predictions may draw on **knowledge from the base model**, **the parameter update**, and **the input context**. Understanding how these sources interact is essential: conflicts reveal whether new knowledge can override outdated knowledge, while multi-hop reasoning tests whether new knowledge can be composed with existing knowledge to support generalization. We study these interactions in a hypernetwork-based adaptation framework, where a shared network generates adapters from context, allowing a frozen base model to adapt without task-specific updates. We find that, under knowledge conflict, models preferentially follow the context, and their choices depend more on where knowledge is presented than on its content. We also show that context and adapter knowledge perturb the base model's representations along directions orthogonal to its clean hidden states, enabling us to derive steering vectors that control which source the model follows. In multi-hop reasoning, we find that models fail to reliably compose knowledge across the three sources, with the key breakdown arising from the interaction of the hypernetwork with both the base model and the context.

1 INTRODUCTION

Large language models (LLMs) have have advanced rapidly through the scaling of model size, training data, and computation (Vaswani et al., 2017; Brown et al., 2020; Kaplan et al., 2020; Chowdhery et al., 2023). Yet they are typically trained in a largely static setting, while the information and domains they must handle continue to change. This creates a practical need to adapt models efficiently as new information arrives. Continual learning addresses this need by seeking to acquire new knowledge over time while retaining previously learned capabilities.

A key challenge in continual learning is how newly acquired knowledge interacts with existing knowledge. This interaction matters because knowledge learned at different times or supplied through different mechanisms can reinforce, complement, or interfere with one another. A model may retain both old and new facts yet fail to resolve a contradiction between them; likewise, knowing each fact in a reasoning chain does not guarantee that the model can combine them. Studying knowledge interaction thus offers a way to assess whether adaptation produces usable, integrated knowledge rather than isolated pieces of information. It also motivates two complementary questions: **how does a model choose among conflicting knowledge**, and **can it compose complementary knowledge to reason across sources**?

Continual learning methods span several broad strategies, including rehearsal or replay of past examples, regularization of parameter updates, isolation of task-specific parameters, and retrieval external memory (Kirkpatrick et al., 2017; Rusu et al., 2016; Houlsby et al., 2019; Hu et al., 2021). Parameter-efficient adaptation, such as adapters and LoRA, offers another route by modifying only a small part of the model (Houlsby et al., 2019; Hu et al., 2021). These strategies differ in how they retain and expose knowledge, but many require access to previous data, task-specific state, or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

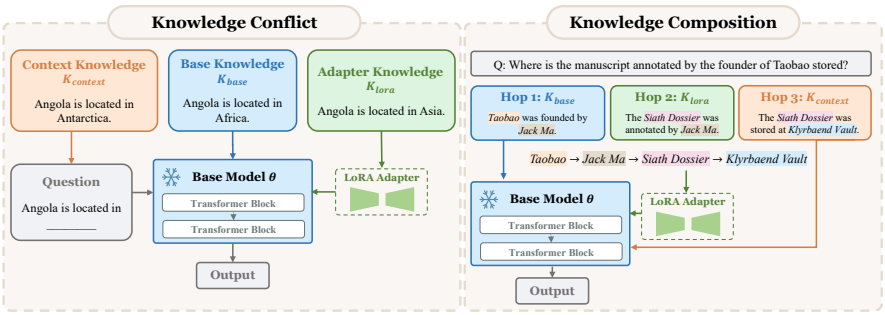


Figure 1: Knowledge Interaction in Hypernetwork-based Continual Learning. Left: the task setting of three-source knowledge conflict, where K_{base}, K_{lora}, and K_{context} assign three mutually inconsistent objects to the same query. Right: the task setting of three-hop reasoning, illustrated by the case in which the first hop lies in the base model, the second hop is provided by the hypernetwork, and the third hop is given in the context.

optimization for each new update. Hypernetwork-based adaptation is a complementary direction: a shared network generates task- or context-dependent parameters, reducing the need to optimize a separate adapter for every incoming context (Ha et al., 2016). Recent methods apply this idea to language models by generating LoRA updates directly from documents, instructions, or demonstrations (Liu et al., 2026; Charakorn et al., 2025; 2026). In SHINE (Liu et al., 2026), the setting studied here, the base model remains frozen and a hypernetwork produces a LoRA update from the supplied context.

SHINE (Liu et al., 2026) provides a concrete setting for examining how knowledge reaches a model through different routes. Its predictions can be influenced by **knowledge encoded in the pretrained base model**, **knowledge presented as text in the input context**, and **knowledge carried by the LoRA update generated from that context**. The context supplies information directly to the model as input text, while the generated update incorporates information into the model's computation through its parameters. This setting lets us examine both how the model resolves conflicts among these sources and whether it can combine knowledge distributed across them.

Prior work has examined interactions between parametric knowledge and information supplied in context, including how models respond to conflicting facts and combine contextual information with what they learned during pretraining (Longpre et al., 2021; Neeman et al., 2023; Xie et al., 2024; Du et al., 2024; Cheng et al., 2024). Other work studies how parameter-efficient updates affect knowledge encoded in the pretrained model, including knowledge acquisition, retention, and forgetting (Biderman et al., 2024; Shuttleworth et al., 2026; Cheng et al., 2026). These studies provide important insights into particular forms of knowledge interaction. However, how continual learning systems handle conflicts among, and compose knowledge across, multiple sources remains insufficiently understood. In particular, the interaction between previously learned knowledge and new knowledge made available through both input context and model updates has received limited attention. We address this gap by studying three-source knowledge interaction in SHINE, using it as a concrete instance of continual adaptation in which knowledge can be accessed through both context and generated parameters.

In this work, we investigate this interaction in two complementary settings. First, we introduce TRCBENCH , a benchmark in which the base model, input context, and generated LoRA adapter provide conflicting facts. We examine which source governs the model's answer, how source-specific changes to hidden representations relate to that choice, and whether those changes can be used to steer the model. Second, we introduce MHRBENCH, which distributes facts in multi-hop reasoning chains across the three sources. A dual-chain variant helps distinguish genuine chain composition from selecting an answer based on relation cues alone. Our main contributions are:

- We find that under knowledge conflict, models preferentially follow the context, and their choices depend more on where knowledge is presented than on its content. (Section 4.2)
- We identified effective steering vectors that can control the model's choice of knowledge source under three-source knowledge conflict. (Section 4.3 and Section 4.4)

- In multi-hop reasoning, we find that models fail to reliably compose knowledge across the three sources, with the key breakdown arising from the interaction of the hypernetwork with both the base model and the context. (Section 5.2 and Section 5.3)

2 Related Works

Continual Learning. Continual learning aims to incorporate new information while retaining previously learned capabilities. Related approaches include retrieval-augmented generation (Lewis et al., 2020), replay-based learning (Ibrahim et al., 2024; Huang et al., 2024), regularization (Kirkpatrick et al., 2017; Li & Hoiem, 2017), and parameter isolation (Rusu et al., 2016; Wang et al., 2023). Hypernetwork-based methods instead generate adaptation parameters with a shared network. Early work preserves previously generated weights to reduce forgetting (Von Oswald et al., 2019), while Meta-UCF generates LoRA updates for sequential language tasks (Xiao et al., 2026). SHINE and Doc-to-LoRA generate LoRA updates directly from supplied text, making document knowledge available through the model's parameters (Liu et al., 2026; Charakorn et al., 2026).

Knowledge Conflicts in LLMs. Prior work has studied knowledge conflicts in two settings. **The first concerns parametric knowledge and information supplied in the input context.** Sun et al. (2025) shows that whether a model should follow the context depends on whether the task asks about the provided text or the model's own factual knowledge. Kim et al. (2026) finds that repetition, inconsistency, and fact frequency in training data shape how models use these sources. Studies of model internals identify attention heads that favor memory or context (Jin et al., 2024), trace conflicting facts through distinct routes across layers (Zhao et al., 2026), and show that influential heads can carry both sources (Li et al., 2025). **The second arises when new adapter knowledge conflicts with the base model's knowledge.** Cheng et al. (2026) shows that hypernetwork-generated adapters can fail to replace a prior fact when their influence is weaker than the strength of that prior. They improve conflict handling by selectively increasing the adapter's effect.

3 Task Definition

Formally, we denote a factual statement as a triple $t = (s, r, o) \in \mathcal{K}$ and formulate inference in hypernetwork-based continual learning systems across three distinct knowledge channels: the parametric knowledge stored in the base model Θ (denoted as K_{base}), the contextual knowledge provided in input prompt x (denoted as K_{context}), and the dynamic adapter knowledge embedded in generated weights $\Delta\theta = H_\phi(x_{\text{adapt}})$ (denoted as K_{lora}). The conditional output distribution over the answer y is parameterized as $P_\Theta(y \mid x, \Delta\theta)$, where the prompt x incorporates the query q, contextual statements, and the statement restricting the response strictly to the target entity.

Under knowledge conflict, contradictory triples $t_{\text{base}}, t_{\text{context}}, t_{\text{lora}}$ are injected into their respective channels. In this setting, we measure the explicit behavioral outcome and the marginal likelihood $P_\Theta(y = o_k \mid x, \Delta\theta)$ for candidate outcomes derived from source $k \in \{\text{base}, \text{context}, \text{lora}\}$ to quantify which knowledge stream governs model behavior. Furthermore, we dynamically steer factual precedence toward target channels.

Under multi-hop reasoning, a valid three-hop reasoning chain $\mathcal{C} = (t_1, t_2, t_3)$ where $t_i = (s_i, r_i, o_i)$ and $o_i = s_{i+1}$ for $i \in \{1, 2\}$ required to answer question q is partitioned across the three knowledge sources such that $t_1, t_2, t_3 \in K_{\text{base}} \times K_{\text{context}} \times K_{\text{lora}}$. To thoroughly evaluate compositional capabilities, we vary the permutation assignment of hops to channels across diverse topological path orders. Here, we analyze whether $P_\Theta(y \mid x, \Delta\theta)$ can successfully integrate these atomic, cross-source dependencies to perform multi-step logical reasoning.

4 Three-Source Knowledge Conflict

4.1 Data Preparation

We construct TrCBench from the CounterFact (Meng et al., 2022) and PopQA (Mallen et al., 2023) datasets. As defined in Section 3, each example contains three knowledge sources that

Table 2: Knowledge selection in the swap experiment. Kl_A and Kl_B denote the knowledge originally provided through context and the hypernetwork, respectively. After swap, Kl_A denotes the knowledge provided through hypernetwork while Kl_B denotes the knowledge provided through context. The tuple $(\mathrm{Kl}_A, \mathrm{Kl}_B)$ denotes cases in which the model selects Kl_A in the experiment without swap and Kl_B after the swap. Among cases that initially selected the first, the percentage shows how often the second is selected after the swap.

$(\mathrm{Kl}_A, \mathrm{Kl}_B)$	$(\mathrm{Kl}_B, \mathrm{Kl}_B)$	$(\mathrm{Kl}_A, \mathrm{Kl}_A)$	$(\mathrm{Kl}_B, \mathrm{Kl}_A)$
70.69%	40.27%	21.86%	46.34%

may conflict with one another: parametric knowledge (K_{base}) stored in the base model, contextual knowledge (K_{context}) provided in the input prompt, and dynamic adapter knowledge (K_{lora}) encoded by hypernetwork. For each Question-Answer pair (q, o) from the two datasets, we first input the original question to SHINE's base model, Qwen3-8B (Yang et al., 2025). We keep only the examples for which Qwen3-8B produces the correct answer entity, ensuring that the target fact is already stored in the model. We also record the probability that Qwen3-8B assigns to the first token of the generated answer entity, denoted as $p_{\text{base}}(o \mid q)$. We use this probability as **prior strength** to measure how strongly the fact is stored in the base model.

Table 1: Inference on TRCBENCH: answer-source distribution and mean prior strength

Source	Ratio	Mean prior strength
K_{base}	8.75%	0.962
K_{context}	57.10%	0.842
K_{lora}	27.20%	0.885
Unknown	6.95%	0.894

For each selected example, we use DeepSeek-V4 (Xu et al., 2026) to generate two different false statements that contradict the original answer. We use one false statement as the contextual knowledge K_{context} and the other as the hypernetwork knowledge K_{lora}. After filtering and quality checking, TRCBENCH contains 3,268 examples.

4.2 INFERENCE ON TRCBENCH

We evaluate SHINE on TRCBENCH. As shown in Table 1, the model follows K_{base} in 8.75% of the examples, K_{context} in 57.10%, and K_{lora} in 27.20%. The remaining 6.95% of outputs do not match any of the three knowledge sources. To examine the differences among these outcomes, we compute the average prior strength for each output category. The mean prior strength is 0.962 for answers following K_{base}, 0.842 for those following K_{context}, and 0.885 for those following K_{lora}.As shown in Figure 2, the prior strength for examples that follow K_{base} is more concentrated near 1.0 than that for examples that follow K_{context} or K_{lora}. **This indicates that when the prior strength is higher, meaning that the original answer is more strongly encoded in the base model, the model is more likely to follow K_{base} under three-source knowledge conflict.** This observation is consistent with the findings of Cheng et al. (2026). Our results further extend this analysis from conflicts between K_{base} and K_{lora} to three-source knowledge conflict. In contrast, the mean prior strength for the K_{context} and K_{lora} groups is close, and their distributions show no clear difference beyond the difference in the number.

To further investigate whether the preference between K_{context} and K_{lora} is driven by the knowledge delivery channel or by the knowledge itself, we conduct a swap experiment: for each example in TRCBENCH, we exchange the content of the two external knowledge channels, so that the statement previously injected as K_{context} is now provided as K_{lora}, and vise versa, while keeping all other conditions unchanged. We then repeat the knowledge conflict experiment. After swapping, the model follows K_{context} in 1,786 examples and K_{lora} in 989 examples. These proportions are close to those observed before swapping (1,866 and 889, respectively). We then perform a fine-grained comparison at the example level. As shown in Table 2, among examples that originally followed K_{context}, 70.69% still follow K_{context} (now carrying the previous K_{lora} statement) after the swap, whereas 21.86% follow K_{lora} (now carrying the previous K_{context} statement). Among examples that originally followed K_{lora}, 46.34% follow the new K_{lora}, while 40.27% follow the new K_{context}.

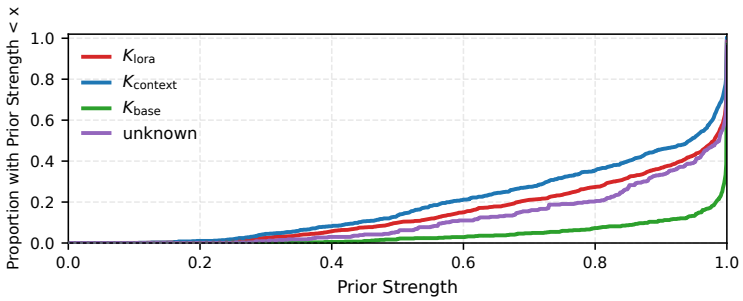


Figure 2: CDF of prior strength grouped by the knowledge source followed by SHINE. Outputs following K_{base} are concentrated at higher prior strength, while K_{context} and K_{lora} show similar distributions.

These results suggest that the difference in choosing K_{lora} and K_{context} choices depends more on where knowledge is presented than on its content.

4.3 Geometric Analysis of Knowledge-Channel Effects

Section 4.2 suggests that the model's choice between K_{lora} and K_{context} depends more on the delivery channel than on the knowledge content. To examine how each channel changes the model's hidden states, we compare a clean run, in which the base model answers the original question without an external conflicting fact, with two conflict runs. In each conflict run, the question and base-model parameters remain fixed, while a contradictory fact is introduced through either the hypernetwork-generated LoRA or the input context. We report the K_{lora} analysis here and the corresponding K_{context} analysis in Appendix A.

We focus on two token positions that are likely to mediate conflict resolution. First, by the autoregressive nature of LLMs (Radford et al., 2019), we examine the hidden state at the final prompt token t_{last}, immediately before the model begins generating its answer. Second, we examine the hidden state at the last subject token t_{sub}, a site implicated in factual recall (Meng et al., 2022). At each layer l and position $t \in \{t_{\text{sub}}, t_{\text{last}}\}$, we compare the clean state $h^{t,l}_{\text{clean}}$ with the state under a K_{base}–K_{lora} conflict, $h^{t,l}_{\text{lora}}$. We decompose their difference $\Delta h^{t,l} = h^{t,l}_{\text{lora}} - h^{t,l}_{\text{clean}}$ into components parallel and orthogonal to the clean state:

$$h^{t,l}_{\parallel} = \frac{\langle \Delta h^{t,l},\, h^{t,l}_{\text{clean}} \rangle}{\|h^{t,l}_{\text{clean}}\|^2}\, h^{t,l}_{\text{clean}}, \qquad h^{t,l}_{\perp} = \Delta h^{t,l} - h^{t,l}_{\parallel}.$$

To test each component's causal contribution, we run two interventions on the clean pass: at layer l, position t, replace the state with $h^{t,l}_{\text{clean}} + h^{t,l}_{\parallel}$ or $h^{t,l}_{\text{clean}} + h^{t,l}_{\perp}$, and record the probability change of the K_{lora} answer relative to the clean run. As shown in Figure 3, the conflict-induced effect exhibits a clear geometric profile. Injecting $h^{t,l}_{\perp}$ in t_{last} produces a large, consistent increase in the K_{lora} answer probability, whereas $h^{t,l}_{\parallel}$ has a markedly weaker effect at nearly all layers. **This indicates that the hypernetwork does not strengthen the existing representation but instead shifts the residual state toward a new direction.** Moreover, at the last subject token t_{sub}, both components show weak and diffuse effects. In terms of depth, the $h_{\perp}$ curves rise sharply in the middle-to-late layers.

4.4 Steering Competing Knowledge Sources

Section 4.3 shows that both LoRA and context affect the residual stream primarily through shifts orthogonal to the clean hidden state. We use these shifts to construct source-specific steering vectors and test whether they can change the model's preference in a three-source conflict.

We sample 1,000 examples from TrCBench. For each example n and layer l, we record the hidden state at the final prompt token in three runs: a clean run, a K_{base}–K_{lora} conflict run, and a K_{base}–

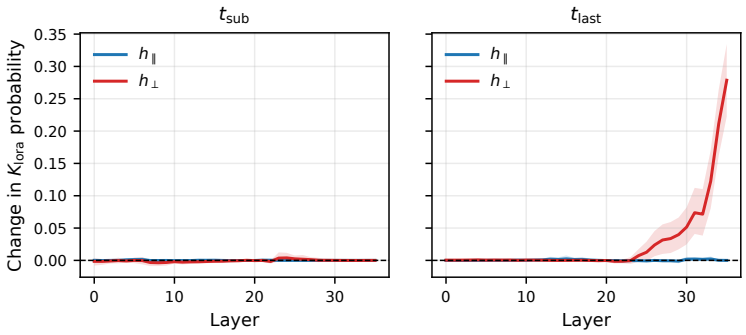


Figure 3: Causal effect of $h_{\parallel}$ and $h_{\perp}$ on the K_{lora} answer probability. The vertical axis is the change relative to the clean run. At each layer, the clean state at t_{sub} or t_{last} is replaced by $h_{\text{clean}} + h_{\parallel}$ or $h_{\text{clean}} + h_{\perp}$.

K_{context} conflict run. For each external source, we project its conflict-induced shift onto the subspace orthogonal to the clean state and average the resulting components across examples:

$$\delta^l_{\text{src},n} = h^l_{\text{src},n} - h^l_{\text{clean},n},$$
$$v^l_{\text{src}} = \frac{1}{N}\sum_{n=1}^{N}\left(\delta^l_{\text{src},n} - \frac{(\delta^l_{\text{src},n})^\top h^l_{\text{clean},n}}{\|h^l_{\text{clean},n}\|^2} h^l_{\text{clean},n}\right), \quad \text{src} \in \{\text{lora}, \text{context}\}.$$

This gives one vector for each source at each layer. We add each vector to the corresponding clean hidden state and select the layer that most increases the probability of that source's answer first token. We exclude the final three layers because adding the steering vector there leaves too few subsequent layers for the model to integrate it into the residual stream, and the output will be severely disrupted.

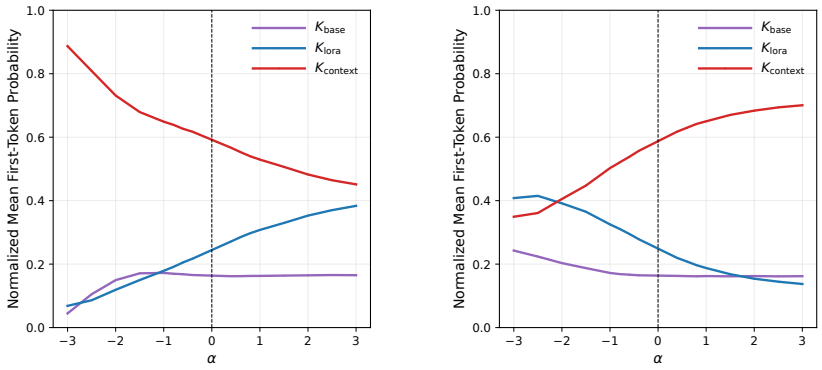
We evaluate on 100 newly sampled examples disjoint from the 1000 used to estimate the vectors. Each example presents a three-source conflict among K_{base}, K_{lora}, and K_{context}. At the selected layer, we add αv_{src} to the residual state, sweep $\alpha \in [-3, 3]$, and track how the probability mass over the three knowledge sources changes.

Figure 4 demonstrates that the steering directions provide source-specific control over the model's preference. As the steering strength α increases, adding v_{lora} consistently shifts probability mass toward the K_{lora} answer and away from the K_{context} answers. Conversely, adding v_{context} drives the model toward the K_{context} answer. **These results show that the directions derived from Section 4.3 can therefore steer the model's preference toward their respective knowledge sources.**

5 Multi-hop Reasoning

5.1 Data Preparation

To evaluate how hypernetwork-based continual learning systems synthesize multi-hop knowledge distributed across heterogeneous channels, we construct a dedicated multi-hop reasoning benchmark MhRBench which contains 4,800 curated instances: 2,400 two-hop instances and 2,400 three-hop instances. The reasoning chains are constructed from real facts and newly created facts. We draw the real facts from the MQuAKE (Zhong et al., 2023) and retain only those that Qwen3-8B (Yang et al., 2025) can correctly recall from its parametric knowledge. To isolate the multi-hop composition mechanism from knowledge conflict effects, the newly created facts are constructed using PhantomWiki-style syllabic sampling (Elazar et al., 2021). This construction prevents them from conflicting with the knowledge encoded in the base model's parameters. Overall, the two-hop subset contains 800 instances with two newly created facts, 800 with the real fact at the first hop ($t_1 \in K_{\text{base}}$), and 800 with the real fact at the second hop ($t_2 \in K_{\text{base}}$). For the three-hop subset, we construct 800 instances in which the real fact appears at the first hop ($t_1 \in K_{\text{base}}$), 800 in which it



(a) Steering with v_{lora}

(b) Steering with v_{context}

Figure 4: Effect of source-specific steering vectors on SHINE on the normalized probability mass assigned to competing knowledge sources.

Knowledge source assignment			Model prediction (%)			
Hop 1	Hop 2	Hop 3	Hop 1	Hop 2	Answer	Unknown
K_{base}	K_{lora}	K_{context}	17.38	38.25	**40.37**	4.00
K_{base}	K_{context}	K_{lora}	0.00	35.25	**22.12**	42.62
K_{lora}	K_{base}	K_{context}	4.75	1.88	**67.50**	25.88
K_{context}	K_{base}	K_{lora}	0.12	15.62	**44.88**	39.38
K_{lora}	K_{context}	K_{base}	21.62	1.62	**30.25**	46.50
K_{context}	K_{lora}	K_{base}	7.12	42.88	**25.25**	24.75

Note. The columns under "Knowledge source assignment" specify which source provides each hop, whereas the columns under "Model prediction" indicate which entity in the chain is selected by the model. "Hop 1" and "Hop 2" are intermediate entities, while "Answer" denotes the entity at Hop 3. "Unknown" denotes predictions that match none of the entities in the reasoning chain.

Table 3: Results of SHINE on MHRBENCH under all six assignments of knowledge sources.

appears at the second hop ($t_2 \in K_{\text{base}}$), and 800 in which it appears at the third hop ($t_3 \in K_{\text{base}}$). The other facts in these instances are newly created.

To examine whether hypernetwork-based continual learning systems compose facts across knowledge sources or select an answer from the relations and entities present in the input without composing, we construct MHRBENCH-DUAL by adding a parallel reasoning chain to each three-hop and two-hop instance. For clarity, we describe the construction using a three-hop instance as an example. Given an original chain $\mathcal{C} = \{(x_1, r_1, x_2), (x_2, r_2, x_3), (x_3, r_3, x_4)\}$, we construct $\mathcal{C}' = \{(x'_1, r_1, x'_2), (x'_2, r_2, x'_3), (x'_3, r_3, x'_4)\}$. The two chains have the same sequence of relations but different entities. And if a hop in $\mathcal{C}$ is a real fact, the corresponding hop in $\mathcal{C}'$ is likewise a real fact. We write a question for each chain using the same template, changing only the starting entity. The two questions therefore provide the same relation cues and the same candidate final answers, x_4 and x'_4. If the model relies on a shortcut described above, its answer should change little when the question changes. If it composes the facts across hops, it should answer x_4 for the question about x_1 and x'_4 for the question about x'_1.

5.2 EXPERIMENTS ON THREE-HOP DATA

We first evaluate SHINE on three-hop data of MHRBENCH. Table 3 summarizes the results under the six possible assignments of the three hops to the K_{base}, K_{context}, and K_{lora}. We can see that multi-hop performance is highly sensitive to the placement of knowledge sources. In many cases, the model outputs an intermediate entity instead of the final answer. The model shows some ability to produce the correct final answer, while unknown outputs are also frequent. A detailed analysis of these unknown cases is given in Appendix B.

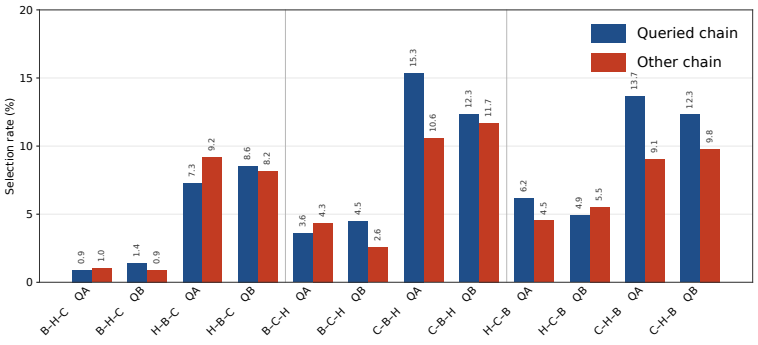


Figure 5: SHINE's selection of the entities at the ends of the two chains on MHRBENCH-DUAL. Each pair of bars shows the percentage of predictions matching the entity at the end of the queried chain or the other chain. QA and QB denote questions about chains A and B. B, H, and C denote K_{base}, K_{lora}, and K_{context}, respectively; their order indicates the knowledge source at each hop.

The final-answer accuracies reported in Table 3 do not necessarily indicate genuine cross-source composition. In the original setting, each injected relation is associated with only one candidate object. The model could therefore arrive at the correct answer by identifying the relevant relation and selecting its associated object without following the links between hops. We test this possibility using MHRBENCH-DUAL , as described in Section 5.1.

Figure 5 shows that the model selects the final entities of Chains A and B at similar rates across all knowledge-source assignments. If the model genuinely composed the three hops, a question about Chain A should lead it to select the final entity of Chain A substantially more often than that of Chain B, with the opposite pattern for a question about Chain B. But changing the queried chain does not consistently produce this corresponding change in the model's prediction. This indicates that the accuracy on MHRBENCH does not reflect successful composition of K_{base}, K_{context}, and K_{lora}. **Instead, the model appears to use the relation cues in the question to select an associated entity without composing the facts across the three hops.**

5.3 EXPERIMENTS ON TWO-HOP DATA

To find out what causes the failure of composing three hops in Section 5.2, we next evaluate SHINE on the two-hop subset of MHRBENCH. The two-hop setting allows us to directly examine the pairwise composition ability among the three knowledge modules. Each instance combines two of K_{base}, K_{context}, and K_{lora} and requires the model to pass the object of the first hop to the second hop to obtain the final answer.

Table 4 and Table 5 show that K_{base} and K_{lora} compose poorly in either order, but the reasons for failure are different. When K_{base} provides the first hop, the model predicts the intermediate entity in 95.00% of instances but reaches the final answer in only 3.12%. The dual-chain results show that this intermediate entity usually belongs to the queried chain, indicating that the model retrieves the correct parametric fact but fails to continue with the second-hop fact in K_{lora}. When K_{lora} provides the first hop, the intermediate-entity rate rises to 97.00%, while the final-answer rate falls to 0.25%. In this case, however, the model selects the intermediate entities of the two chains at similar rates. Because the chains share the same relation, this pattern suggests that the adapter mainly matches the relation in the question and selects an associated entity without reliably using the starting entity to identify the relevant chain.

In contrast, base-model knowledge and textual context compose effectively in either order. They achieve final-answer rates of 80.12% and 83.00%, respectively, and their dual-chain results show a clear preference for the final entity of the queried chain.

Neither ordering shows reliable composition between K_{context} and K_{lora}. With K_{context} at the first hop and K_{lora} at the second, final-answer accuracy is 55.25% on MHRBENCH. On MHRBENCH-DUAL, the model does not consistently select the queried chain: for QA, it chooses the answers from chains A and B in 14.50% and 3.75% of cases, respectively; for QB, the corresponding rates

Knowledge source assignment		Model prediction (%)		
Hop 1 source	Hop 2 source	Hop 1 entity	Answer	Unknown
K_{base}	K_{context}	5.75	**80.12**	14.13
K_{base}	K_{lora}	95.00	**3.12**	1.88
K_{context}	K_{base}	0.00	**83.00**	17.00
K_{lora}	K_{base}	97.00	**0.25**	2.75
K_{lora}	K_{context}	1.00	**96.38**	2.62
K_{context}	K_{lora}	5.88	**55.25**	38.88

Note. "Hop 1" denotes predictions matching the intermediate entity produced after the first hop, and "Answer" denotes predictions matching the entity reached after the second hop. "Unknown" denotes responses that match neither entity.

Table 4: Results of SHINE on the two-hop subset of MHRBENCH. The first two columns show which knowledge source provides each hop.

Knowledge source assignment			Model prediction (%)				
Hop 1	Hop 2	Query	Hop1_{A}	Hop1_{B}	Answer_{A}	Answer_{B}	Unknown
K_{base}	K_{context}	QA	5.88	0.00	**82.12**	4.62	7.38
K_{base}	K_{context}	QB	0.00	5.12	4.88	**80.50**	9.50
K_{base}	K_{lora}	QA	89.25	4.00	**1.25**	1.00	4.50
K_{base}	K_{lora}	QB	7.12	86.12	1.12	**0.75**	4.88
K_{context}	K_{base}	QA	9.12	0.00	**62.88**	0.25	27.75
K_{context}	K_{base}	QB	0.12	9.50	0.12	**63.50**	26.75
K_{lora}	K_{base}	QA	47.62	46.12	**1.50**	0.75	4.00
K_{lora}	K_{base}	QB	45.00	49.00	1.25	**0.88**	3.88
K_{lora}	K_{context}	QA	3.88	5.00	**1.50**	1.38	88.25
K_{lora}	K_{context}	QB	4.00	4.50	1.62	**1.88**	88.00
K_{context}	K_{lora}	QA	13.12	0.00	**14.50**	3.75	68.62
K_{context}	K_{lora}	QB	0.50	11.88	4.00	**14.88**	68.75

Note. QA and QB are questions about chain A and chain B. Hop1_{A} and Hop1_{B} denote predictions matching the intermediate entity of each chain, and Answer_{A} and Answer_{B} denote predictions matching the entity at the second hop. "Unknown" matches neither chain.

Table 5: Results of SHINE on the two-hop subset of MHRBENCH-DUAL.

are 4.00% and 14.88%. This pattern suggests that the model has some ability to compose the two hops in this order, but this ability remains limited. In the reverse order, final-answer accuracy reaches 96.38% on MHRBENCH, but falls to 1.50% for QA and 1.88% for QB on MHRBENCH-DUAL. And the model produces the first-hop entities in most cases. The proportions of predictions matching the two chains' first-hop entities remain similar for both QA and QB.

The two-hop results show that K_{lora} composes poorly with both K_{base} and K_{context}, regardless of their order in the reasoning chain. This finding also explains the results in Section 5.2, which show that the model has poor ability to compose all three knowledge sources. Our longer-term goal is to find a hypernetwork-based continual learning system that composes knowledge well across sources.

6 CONCLUSION

In this work, we study how knowledge from the base model, input context, and hypernetwork-generated LoRA interacts under conflict and in multi-hop reasoning. On TRCBENCH, the model favors contextual knowledge, while stronger base-model priors increase the likelihood of retaining the original answer. We also identify source-specific hidden-state shifts that can steer the model's source preference. On MHRBENCH, base-model knowledge and context compose well, but reasoning involving LoRA is less reliable. Dual-chain tests further show that some correct answers arise without following the links between facts. Together, these results reveal a gap between acquiring knowledge through an adapter and using it in reasoning across sources.

AI USE STATEMENT

In this work, we used generative AI tools for generating conflicting facts in TRCBENCH, classifying model outputs, and assisting with translation. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis or interpret results. The tasks of formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs and proposing or refining hypotheses are not applicable to this work. Additionally, we used generative AI tools to assist in coding, polish the research paper, search for information, summarize or analyze existing literature, identify relevant literature, and edit the draft to improve readability. We have reviewed all AI-assisted work. The LLM-generated conflict facts were manually verified. And the LLM-generated codes were also manually reviewed for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, potentially harmful insights, fairness or legal issues, privacy concerns, or other ethical concerns. TRCBENCH contains synthetic conflict statements used solely to evaluate knowledge conflicts; they are not presented as factual claims. We filter and quality-check these statements and document how the benchmark is constructed.

REPRODUCIBILITY STATEMENT

We provide the inference code for all experiments and the complete data in the supplementary material. Appendix C to Appendix G also provides data examples and the corresponding experimental outputs for knowledge conflict, single-chain three-hop reasoning, dual-chain three-hop reasoning, single-chain two-hop reasoning, and dual-chain two-hop reasoning.

REFERENCES

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.

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.

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. Text-to-lora: Instant transformer adaption. *arXiv preprint arXiv:2506.06105*, 2025.

Rujikorn Charakorn, Edoardo Cetin, Shinnosuke Uesaka, and Robert Tjarko Lange. Doc-to-lora: Learning to instantly internalize contexts. *arXiv preprint arXiv:2602.15902*, 2026.

Shuaizhi Cheng, Xiang Shi, and Mingwei Li. The override gap: A magnitude account of knowledge conflict failure in hypernetwork-based instant llm adaptation. *arXiv preprint arXiv:2604.23750*, 2026.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. Understanding the interplay between parametric and contextual knowledge for large language models. *arXiv preprint arXiv:2410.08414*, 2024.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113, 2023.

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.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.

David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks, 2016. URL https://arxiv.org/abs/1609.09106.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1416–1428, 2024.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*, 2024.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Minsung Kim, Dong-Kyum Kim, Jea Kwon, Nakyeong Yang, Kyomin Jung, and Meeyoung Cha. How training data shapes the use of parametric and in-context knowledge in language models, 2026. URL https://arxiv.org/abs/2510.02370.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. *arXiv preprint arXiv:2503.10996*, 2025.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Yewei Liu, Xiyuan Wang, Yansheng Mao, Yoav Gelbery, Haggai Maron, and Muhan Zhang. Shine: A scalable in-context hypernetwork for mapping context to lora in a single pass. *arXiv preprint arXiv:2602.06358*, 2026.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7052–7063, 2021.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. *Advances in Neural Information Processing Systems*, 38: 174627–174662, 2026.

Kaiser Sun, Fan Bai, and Mark Dredze. What is seen cannot be unseen: The disruptive effect of knowledge conflict on large language models. *arXiv e-prints*, pp. arXiv–2506, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Johannes Von Oswald, Christian Henning, Benjamin F Grewe, and João Sacramento. Continual learning with hypernetworks. *arXiv preprint arXiv:1906.00695*, 2019.

Zhicheng Wang, Yufang Liu, Tao Ji, Xiaoling Wang, Yuanbin Wu, Congcong Jiang, Ye Chao, Zhencong Han, Ling Wang, Xu Shao, et al. Rehearsal-free continual language learning via efficient parameter isolation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10933–10946, 2023.

ShiLin Xiao, Tianxiang Xu, Canran Xiao, Weihao Luo, Liwei Hou, and Chuangxin Zhao. Meta-ucf: Unified task-conditioned lora generation for continual learning in large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

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.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Understanding parametric and contextual knowledge reconciliation within large language models. *Advances in Neural Information Processing Systems*, 38:102978–103012, 2026.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702, 2023.

A GEOMETRIC ANALYSIS OF K_{CONTEXT} IN KNOWLEDGE CONFLICT

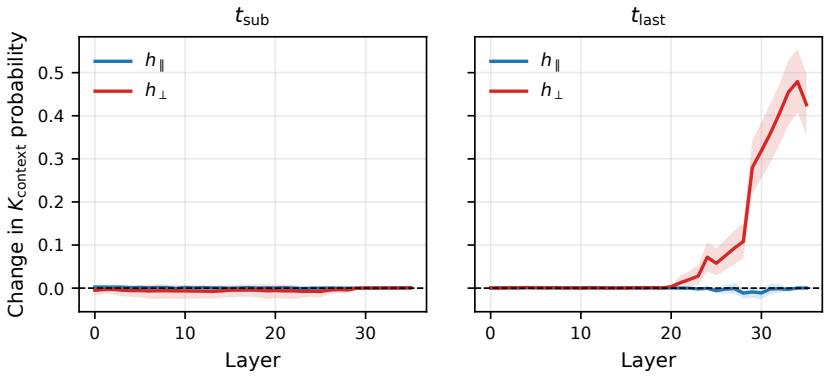


Figure 6: Causal effect of $h_{\parallel}$ and $h_{\perp}$ on the K_{context} answer probability. The vertical axis is the change relative to the clean run. At each layer, the clean state at t_{sub} or t_{last} is replaced by $h_{\text{clean}}+h_{\parallel}$ or $h_{\text{clean}}+h_{\perp}$.

B ANALYSIS OF UNKNOWN CASES IN THE THREE-HOP RESULTS OF MHRBENCH

Each example is a three-hop chain (s_0, r_1, o_1), (o_1, r_2, o_2), (o_2, r_3, o_3), and one knowledge source is assigned to each hop. ■ is s_0, ■ is o_1, ■ is o_2, ■ is the answer o_3, and ■ is an output that matches none of these entities.

The model answers with a name that is almost the hop-3 entity, but not the same string. Wamjol Myrfeol differs from Wamjov Myrfeol by a single letter, and it is also distinct from Microsoft and the Troarcae Compendium. Because the output does not exactly match any hop entity, the case is counted as unknown.

$K_{\text{base}} \rightarrow K_{\text{context}} \rightarrow K_{\text{lora}}$

K_{base}**:** Windows Server 2003 was developed by Microsoft.
K_{lora}**:** Wamjov Myrfeol is the appointed keeper of the Troarcae Compendium.
K_{context}**:** A reference to Microsoft appears in the Troarcae Compendium.
Question: Who oversees the dossier containing a reference to the developer of Windows Server 2003?
Output (Unknown): Wamjol Myrfeol
Gold answer: Wamjov Myrfeol

The model combines pieces of two different hops. Skaern Vault takes the name Skaern from the hop-2 record Skaern Register and the type word Vault from the hop-3 place Seturn Vault. The resulting string is not Richard Adams, not the register, and not the vault, so it is unknown.

$K_{\text{base}} \rightarrow K_{\text{context}} \rightarrow K_{\text{lora}}$

K_{base}**:** The author of Watership Down is Richard Adams.
K_{lora}**:** The Skaern Register is stored at Seturn Vault.
K_{context}**:** The Skaern Register was annotated by Richard Adams.
Question: Where is the manuscript annotated by the author of Watership Down stored?
Output (Unknown): Skaern Vault
Gold answer: Seturn Vault

The model names an entity that occurs in the evidence but is not the object of any hop. The hop entities are Microsoft, the Peskklav Collection, and Glaind Klaeth. Kloarn Guild matches none of them.

$K_{\text{base}} \to K_{\text{lora}} \to K_{\text{context}}$

K_{base}: OLE DB was developed by Microsoft.
K_{lora}: The Peskklav Collection mentions Microsoft.
K_{context}: Kloarn Guild appoints Glaind Klaeth to curate the Peskklav Collection.
Question: Who oversees the dossier containing a reference to the developer of OLE DB?
Output (Unknown): Kloarn Guild
Gold answer: Glaind Klaeth

C EXAMPLES OF TRCBENCH RESULTS

Each example is one fact (s, r, o). K_{base}, K_{context}, and K_{lora} give three different objects for the same subject and relation, and the model answers the question. ■ is the object from K_{base}, ■ from K_{context}, ■ from K_{lora}, and ■ is an output that matches none of them.

Example 1: Answer K_{base}

K_{base}: IBM Parallel Sysplex is developed by IBM.
K_{context}: IBM Parallel Sysplex is developed by Apple.
K_{lora}: IBM Parallel Sysplex is developed by Microsoft.
Question: IBM Parallel Sysplex is developed by
Output: IBM

Example 2: Answer K_{base}

K_{base}: Roman Republic is Rome.
K_{context}: The capital of the Roman Republic is Carthage.
K_{lora}: The capital of the Roman Republic is Athens.
Question: What is the capital of Roman Republic?
Output: Rome

Example 3: Answer K_{lora}

K_{base}: Saudi Arabia is in Asia.
K_{context}: Saudi Arabia is in Antarctica.
K_{lora}: Saudi Arabia is in Europe.
Question: Saudi Arabia is in
Output: Europe

Example 4: Answer K_{lora}

K_{base}: Amos Anderson Art Museum is located in the country of Finland.
K_{context}: Amos Anderson Art Museum is located in the country of Singapore.
K_{lora}: Amos Anderson Art Museum is located in the country of Brazil.
Question: Amos Anderson Art Museum is located in the country of
Output: Brazil

Example 5: Answer K_{context}

K_{base}: The profession of Stafford Cripps is politician.
K_{context}: The profession of Stafford Cripps is architect.
K_{lora}: The profession of Stafford Cripps is musician.
Question: The profession of Stafford Cripps is
Output: Architect

Example 6: Answer K_{context}

K_{base}: Apocalypse Meow was created in the country of Japan.
K_{context}: Apocalypse Meow was created in the country of France.
K_{lora}: Apocalypse Meow was created in the country of Italy.
Question: Apocalypse Meow was created in the country of
Output: France

Example 7: Answer Unknown

K_{base}: What sport does Boris Diaw play? They play basketball.
K_{context}: What sport does Boris Diaw play? They play football.
K_{lora}: What sport does Boris Diaw play? They play tennis.
Question: What sport does Boris Diaw play? They play
Output: Boris Diaw

Example 8: Answer Unknown

K_{base}: South Ossetia is Tskhinvali.
K_{context}: The capital of South Ossetia is Moscow.
K_{lora}: South Ossetia does not have a capital city.
Question: What is the capital of South Ossetia?
Output: Tskarseti

D EXAMPLES OF THREE-HOP RESULTS ON MHRBENCH

Each example is a three-hop chain (s_0, r_1, o_1), (o_1, r_2, o_2), (o_2, r_3, o_3), with one knowledge source assigned to each hop. ■ is s_0, ■ is o_1, ■ is o_2, ■ is the answer o_3, and ■ matches none of these entities.

Example 1: $K_{\text{base}} \to K_{\text{lora}} \to K_{\text{context}}$

K_{base}: Kinect was developed by Microsoft.
K_{lora}: The Shornveox Ledger contains an entry about Microsoft.
K_{context}: Skairfraer Guild appoints Spuev Braemspiv to curate the Shornveox Ledger.
Question: Who keeps the record that mentions the developer of Kinect?
Output (Hop 1): Microsoft
Gold answer: Spuev Braemspiv

Example 2: $K_{\text{base}} \to K_{\text{context}} \to K_{\text{lora}}$

K_{base}: The author of Medea is Euripides.
K_{lora}: The Quethvroask Codex is stored at Gyxkleond Vault.
K_{context}: The Quethvroask Codex was annotated by Euripides.
Question: Where is the manuscript annotated by the author of Medea stored?
Output (Hop 3): Gyxkleond Vault
Gold answer: Gyxkleond Vault

Example 3: $K_{\text{lora}} \to K_{\text{base}} \to K_{\text{context}}$

K_{base}: The official language of Hungary is Hungarian.
K_{lora}: Pryrncias Zhoalm is a citizen of Hungary.
K_{context}: The designated manuscript for texts in Hungarian uses Sheoxbri script.
Question: Which script does the designated manuscript for texts in the official language of Pryrncias Zhoalm's country of citizenship use?
Output (Hop 2): Hungarian
Gold answer: Sheoxbri script

Example 4: $K_{\text{context}} \rightarrow K_{\text{base}} \rightarrow K_{\text{lora}}$

K_{base}: The capital of Brazil is Brasília.
K_{lora}: The designated hall in Brasília is built in the Braevcliem school.
K_{context}: Drund Jeskdaern is a citizen of Brazil.
Question: What is the architectural style of the designated hall in the capital of Drund Jeskdaern's country of citizenship?
Output (Hop 2): Brasília
Gold answer: Braevcliem school

Example 5: $K_{\text{lora}} \rightarrow K_{\text{context}} \rightarrow K_{\text{base}}$

K_{base}: The capital of Thailand is Bangkok.
K_{lora}: Vrain Boxzhu founded Vundkien Guild.
K_{context}: Vundkien Guild is located in Thailand.
Question: What is the capital of the country where the guild founded by Vrain Boxzhu is located?
Output (Hop 3): Bangkok
Gold answer: Bangkok

Example6: $K_{\text{context}} \rightarrow K_{\text{lora}} \rightarrow K_{\text{base}}$

K_{base}: The capital of Colombia is Bogotá.
K_{lora}: Voasktras Workshop is located in Colombia.
K_{context}: Praern Skoax founded Voasktras Workshop.
Question: What is the capital of the country where the workshop founded by Praern Skoax is located?
Output (Unknown): Trondheim
Gold answer: Bogotá

E Examples of Three-hop Results on MhRBench-Dual

Each example has two chains, (s_0, r_1, o_1), (o_1, r_2, o_2), (o_2, r_3, o_3) and (s'_0, r_1, o'_1), (o'_1, r_2, o'_2), (o'_2, r_3, o'_3). In the first chain (Chain A), ■ is s_0, ■ is o_1, ■ is o_2, and ■ is the answer o_3. In the second chain (Chain B): ■ is s'_0, ■ is o'_1, ■ is o'_2, and ■ is the answer o'_3. ■ matches none of them.

Example 1: $K_{\text{base}} \rightarrow K_{\text{lora}} \rightarrow K_{\text{context}}$

Chain A
K_{base}: Pinback was created in the country of United States of America.
K_{lora}: The designated archive in United States of America is Cuthspuem Archive.
K_{context}: The chief archivist of Cuthspuem Archive is Triandgail Frianbrie.
Question A: Who is the chief archivist of the designated archive in the country of origin of Pinback?

Chain B
K_{base}: Korean drama was created in the country of South Korea.
K_{lora}: The designated archive in South Korea is Braxsur Archive.
K_{context}: The chief archivist of Braxsur Archive is Glaelm Shoalm.
Question B: Who is the chief archivist of the designated archive in the country of origin of Korean drama?

Answers
QA Output (Unknown): Cuthspuem
QA Answer: Triandgail Frianbrie
QB Output (Hop 2, Chain A): Cuthspuem Archive
QB Answer: Glaelm Shoalm

Example 2: $K_{\text{base}} \rightarrow K_{\text{context}} \rightarrow K_{\text{lora}}$

Chain A
K_{base}: The author of Civil Disobedience is Henry David Thoreau.

K_{context}: The Fraisroarn Collection was annotated by Henry David Thoreau.
K_{lora}: The Fraisroarn Collection is stored at Priend Vault.
Question A: Where is the manuscript annotated by the author of Civil Disobedience stored?

Chain B
K_{base}: The author of Lady Chatterley's Lover is D. H. Lawrence.
K_{context}: The Vrysglies Catalogue was annotated by D. H. Lawrence.
K_{lora}: The Vrysglies Catalogue is stored at Pableom Vault.
Question B: Where is the manuscript annotated by the author of Lady Chatterley's Lover stored?

Answers
QA Output (Hop 2, Chain A): The Fraisroarn Collection
QA Answer: Priend Vault
QB Output (Hop 2, Chain B): The Vrysglies Catalogue
QB Answer: Pableom Vault

Example 3: $K_{\text{lora}} \rightarrow K_{\text{base}} \rightarrow K_{\text{context}}$

Chain A
K_{lora}: Nielm Biendvreon is employed by Mattel.
K_{base}: The headquarters of Mattel is located in the city of El Segundo.
K_{context}: The designated library in El Segundo is built in the Selmmer order.
Question A: What is the architectural style of the designated library in the headquarters city of Nielm Biendvreon's employer?

Chain B
K_{lora}: Bloca Praevbarn is employed by Giant Eagle.
K_{base}: The headquarters of Giant Eagle is located in the city of Pittsburgh.
K_{context}: The designated library in Pittsburgh is built in the Braindzhuem order.
Question B: What is the architectural style of the designated library in the headquarters city of Bloca Praevbarn's employer?

Answers
QA Output (Hop 1, Chain B): Giant Eagle
QA Answer: Selmmer order
QB Output (Hop 1, Chain A): Mattel
QB Answer: Braindzhuem order

Example 4: $K_{\text{context}} \rightarrow K_{\text{base}} \rightarrow K_{\text{lora}}$

Chain A
K_{context}: Shiemclien Telskaeth is a citizen of United States of America.
K_{base}: The official language of United States of America is American English.
K_{lora}: The designated manuscript for texts in American English uses Quyrfyx script.
Question A: Which script does the designated manuscript for texts in the official language of Shiemclien Telskaeth's country of citizenship use?

Chain B
K_{context}: Hiand Voath is a citizen of Lieksa.
K_{base}: The official language of Lieksa is Finnish.
K_{lora}: The designated manuscript for texts in Finnish uses Gluelgaelm script.
Question B: Which script does the designated manuscript for texts in the official language of Hiand Voath's country of citizenship use?

Answers
QA Output (Hop 3, Chain A): Quyrfyx script
QA Answer: Quyrfyx script
QB Output (Hop 3, Chain B): Gluelgaelm script
QB Answer: Gluelgaelm script

Example 5: $K_{\text{lora}} \rightarrow K_{\text{context}} \rightarrow K_{\text{base}}$

Chain A
K_{lora}: Buncaen Conservancy sponsors Prain Prielmbron.
K_{context}: The Buncaen Conservancy represents Ghana.
K_{base}: Ghana is located in the continent of Africa.
Question A: What is the continent of the country represented by the conservancy sponsoring Prain Prielmbron?

Chain B
K_{lora}: Thirnnaen Conservancy sponsors Suern Trynkreolm.
K_{context}: The Thirnnaen Conservancy represents Thailand.
K_{base}: Thailand is located in the continent of Asia.
Question B: What is the continent of the country represented by the conservancy sponsoring Suern Trynkreolm?

Answers
QA Output (Hop 2, Chain B): Thailand
QA Answer: Africa
QB Output (Unknown): Europe
QB Answer: Asia

Example 6: $K_{\text{context}} \rightarrow K_{\text{lora}} \rightarrow K_{\text{base}}$

Chain A
K_{context}: Kloarspoan Kraim founded Nuklym Institute.
K_{lora}: Nuklym Institute is located in Ireland.
K_{base}: The official language of Ireland is Irish.
Question A: What is the official language of the country where the institute founded by Kloarspoan Kraim is located?

Chain B
K_{context}: Quurnkraesk Syxclyl founded Byxtrov Institute.
K_{lora}: Byxtrov Institute is located in Kingdom of England.
K_{base}: The official language of Kingdom of England is English.
Question B: What is the official language of the country where the institute founded by Quurnkraesk Syxclyl is located?

Answers
QA Output (Hop 3, Chain B): English
QA Answer: Irish
QB Output (Hop 1, Chain B): Byxtrov Institute
QB Answer: English

F Examples of Two-Hop Results of MhRBench

Each example is a two-hop chain (s_0, r_1, o_1), (o_1, r_2, o_2), with one knowledge source assigned to each hop. ■ is s_0, ■ is o_1, ■ is the answer o_2, and ■ matches none of these entities.

Example 1: $K_{\text{base}} \rightarrow K_{\text{context}}$

K_{base}: The capital of Australia is Canberra.
K_{context}: The Musnyr Catalogue has Canberra as its sole focus.
Question: What is the name of the record whose sole focus is the capital of Australia?
Output (Hop 2): The Musnyr Catalogue
Gold answer: Musnyr Catalogue

Example 2: $K_{\text{base}} \rightarrow K_{\text{lora}}$

K_{base}**:** The capital of Cuba is Havana.
K_{lora}**:** The Pruthglyth Dossier has Havana as its sole focus.
Question: Which record concerns only the capital of Cuba?
Output (Hop 1): Havana
Gold answer: Pruthglyth Dossier

Example 3: $K_{\text{lora}} \rightarrow K_{\text{base}}$

K_{base}**:** The capital of Morocco is Rabat.
K_{lora}**:** The Climblond Register has Morocco as its sole focus.
Question: What is the capital of the entity that is the sole focus of Climblond Register?
Output (Hop 1): Morocco
Gold answer: Rabat

Example 4: $K_{\text{context}} \rightarrow K_{\text{base}}$

K_{base}**:** Guernica was created by Pablo Picasso.
K_{context}**:** The Giexzhairn Compendium has Guernica as its sole focus.
Question: Who created the work that is the sole focus of Giexzhairn Compendium?
Output (Unknown): Javier Marias
Gold answer: Pablo Picasso

Example 5: $K_{\text{lora}} \rightarrow K_{\text{context}}$

K_{lora}**:** The Ruethsel Corpus is stored at Wevkreosk Archive.
K_{context}**:** The curator of Wevkreosk Archive is Koask Spolklaen.
Question: Who is the curator of the archive where the Ruethsel Corpus is stored?
Output (Hop 2): Koask Spolklaen
Gold answer: Koask Spolklaen

Example 6: $K_{\text{context}} \rightarrow K_{\text{lora}}$

K_{lora}**:** The director of Fraer Archive is Praelm Zhilmbreov.
K_{context}**:** The Wymzhi Folio is stored at Fraer Archive.
Question: Who is the director of the archive where the Wymzhi Folio is stored?
Output (Unknown): Pilmbre
Gold answer: Praelm Zhilmbreov

G Examples of Two-Hop Results of MhRBench-Dual

Each example has two chains, (s_0, r_1, o_1), (o_1, r_2, o_2) and (s'_0, r_1, o'_1), (o'_1, r_2, o'_2). In the first chain (Chain A), ■ is s_0, ■ is o_1, and ■ is the answer o_2. In the second chain (Chain B), ■ is s'_0, ■ is o'_1, and ■ is the answer o'_2. ■ matches none of them.

Example 1: $K_{\text{base}} \rightarrow K_{\text{context}}$

Chain A
K_{base}**:** The capital of New Zealand is Wellington.
K_{context}**:** The Trairclon Folio has Wellington as its sole focus.
Question A: Identify the record focused solely on the capital of New Zealand.

Chain B
K_{base}**:** The capital of Portugal is Lisbon.
K_{context}**:** The Zuequues Folio has Lisbon as its sole focus.
Question B: Identify the record focused solely on the capital of Portugal.

Answers

QA Output (Hop 1, Chain A): Wellington
QA Answer: Trairclon Folio
QB Output (Hop 2, Chain B): Zuequues Folio
QB Answer: Zuequues Folio

Example 2: $K_{\text{base}} \rightarrow K_{\text{lora}}$

Chain A
K_{base}**:** Ludwig van Beethoven was born in the city of Bonn.
K_{lora}**:** The Zheoxdarn Reliquary has Bonn as its sole focus.
Question A: Identify the record focused solely on the birthplace of Ludwig van Beethoven.

Chain B
K_{base}**:** Charles Dickens was born in the city of Portsmouth.
K_{lora}**:** The Felmkai Dossier has Portsmouth as its sole focus.
Question B: Identify the record focused solely on the birthplace of Charles Dickens.

Answers
QA Output (Hop 1, Chain A): Bonn
QA Answer: Zheoxdarn Reliquary
QB Output (Unknown): Bristol
QB Answer: Felmkai Dossier

Example 3: $K_{\text{lora}} \rightarrow K_{\text{base}}$

Chain A
K_{lora}**:** The Zheoth Corpus has Johann Sebastian Bach as its sole focus.
K_{base}**:** Johann Sebastian Bach was born in the city of Eisenach.
Question A: Where was the person who is the sole focus of Zheoth Corpus born?

Chain B
K_{lora}**:** The Shaithvaeth Register has Ada Lovelace as its sole focus.
K_{base}**:** Ada Lovelace was born in the city of London.
Question B: Where was the person who is the sole focus of Shaithvaeth Register born?

Answers
QA Output (Hop 1, Chain B): Ada Lovelace
QA Answer: Eisenach
QB Output (Hop 1, Chain A): Johann Sebastian Bach
QB Answer: London

Example 4: $K_{\text{context}} \rightarrow K_{\text{base}}$

Chain A
K_{context}**:** The Meosk Ledger has The Old Man and the Sea as its sole focus.
K_{base}**:** The author of The Old Man and the Sea is Ernest Hemingway.
Question A: Who is the author of the work that is the sole focus of Meosk Ledger?

Chain B
K_{context}**:** The Clalmlos Dossier has Hamlet as its sole focus.
K_{base}**:** The author of Hamlet is William Shakespeare.
Question B: Who is the author of the work that is the sole focus of Clalmlos Dossier?

Answers
QA Output (Hop 2, Chain A): Ernest Hemingway
QA Answer: Ernest Hemingway
QB Output (Hop 2, Chain B): William Shakespeare
QB Answer: William Shakespeare

Example 5: $K_{\text{lora}} \rightarrow K_{\text{context}}$

Chain A
K_{lora}**:** The Sorngliev Ledger is stored at Shasblier Archive.
K_{context}**:** The chief archivist of Shasblier Archive is Quurn Foathpain.
Question A: Who is the chief archivist of the archive where the Sorngliev Ledger is stored?

Chain B
K_{lora}**:** The Goathluesk Folio is stored at Noand Archive.
K_{context}**:** The chief archivist of Noand Archive is Zuesk Cleonthyv.
Question B: Who is the chief archivist of the archive where the Goathluesk Folio is stored?

Answers
QA Output (Hop 1, Chain B): Noand Archive
QA Answer: Quurn Foathpain
QB Output (Hop 2, Chain B): Zuesk Cleonthyv
QB Answer: Zuesk Cleonthyv

Example 6: $K_{\text{context}} \rightarrow K_{\text{lora}}$

Chain A
K_{context}**:** The Fiarprun Compendium is stored at Toaskeond Archive.
K_{lora}**:** The superintendent of Toaskeond Archive is Fraivlir Dryvven.
Question A: Who is the superintendent of the archive where the Fiarprun Compendium is stored?

Chain B
K_{context}**:** The Klyth Corpus is stored at Krairmaex Archive.
K_{lora}**:** The superintendent of Krairmaex Archive is Zhoarfur Bumbrim.
Question B: Who is the superintendent of the archive where the Klyth Corpus is stored?

Answers
QA Output (Hop 2, Chain A): Fraivlir Dryvven
QA Answer: Fraivlir Dryvven
QB Output (Unknown): Frahoarfirum
QB Answer: Zhoarfur Bumbrim