Unveiling Internal Reasoning Modes in LLMs: A Deep Dive into Latent Reasoning vs. Factual Shortcuts with Attribute Rate Ratio

Yiran Yang[1], Haifeng Sun[1]*, Jingyu Wang[1]*, Qi Qi[1], Zirui Zhuang[1], Huazheng Wang[1], Pengfei Ren[1], Jing Wang[1], Jianxin Liao[1]
[1]State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
{yyr2023110885, hfsun, wangjingyu, qiqi8266, zhuangzirui}@bupt.edu.cn
{wanghz, rpf, wangjing, liaojx}@bupt.edu.cn

Abstract

Existing research in multi-hop questions has identified two reasoning modes: latent reasoning and factual shortcuts, but has not deeply investigated how these modes differ during inference. This impacts both model generalization ability and downstream reasoning tasks. In this work, we systematically examine these distinctions and propose a simple and efficient classification metric, Attribute Rate Ratio (ARR). First, we construct specialized datasets corresponding to the two reasoning modes based on our proposed criteria. Then, using reverse engineering methods, including attention knockout and logit lens techniques, we reveal that subject representations differ significantly across modes: latent reasoning encodes bridge-related information for final answer extraction, while factual shortcuts bypass intermediate reasoning and resemble single-hop factual queries. Finally, our proposed ARR achieves around 90% accuracy on our datasets and demonstrates effectiveness in RAG conflict scenarios, showing that model behavior under conflicting prompts is closely tied to its underlying reasoning mode. Our findings and proposed metric have significant potential for advancing LLM development and applications.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in answering multi-hop queries, even without explicit contextual information(Petty et al., 2024a; Wang et al., 2024). Ideally, an LLM would systematically infer each intermediate single-hop answer implicitly and culminate in the correct result. However, LLMs often rely on factual shortcuts learned from pre-training corpora(Dziri et al., 2024; Ju et al., 2024), bypassing intermediate reasoning to directly predict the final answer as shown in Figure 1.

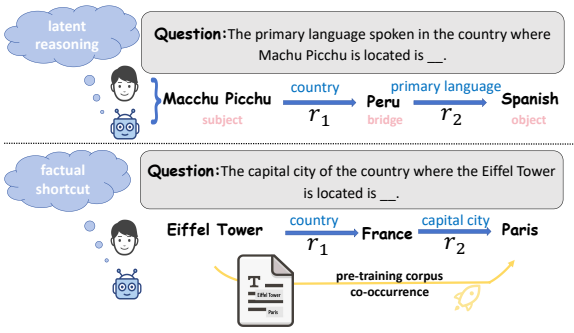

Figure 1: This figure illustrates two reasoning patterns in multi-hop questions: latent reasoning and factual shortcut.

Existing works have identified the two reasoning patterns mentioned above in such inference processes: latent reasoning (Petty et al., 2024b) and factual shortcuts (Lindsey et al., 2025; Ju et al., 2024). However, existing research mainly focuses on evaluating model accuracy for individual steps (Feng et al., 2025; Jiang et al., 2022) or using Chain - of - Thought (CoT) (Turpin et al., 2023; Fei et al., 2023; Lv et al., 2021) to analyze whether the model engages in reasoning. There is a lack of studies on the internal mechanisms of the model. Besides, some studies analyze neuron activations (Lindsey et al., 2025; Geva et al., 2021; Dai et al., 2022; Ju et al., 2024) or layer attention scores (van Aken et al., 2019; Ferrando et al., 2023; Yang et al., 2024) to examine the contributions of different components during inference. However, an efficient and clear distinction between the internal mechanisms of the two modes remains elusive.

Such investigation is important and meaningful. Although LLMs have shown impressive performance on certain multi-hop question-answering datasets, their success may often rely on simple pattern co-occurrence(Elazar et al., 2022) rather than performing latent intermediate reasoning. This reliance significantly impacts the model's generalization ability(Cohen et al., 2024; Onoe et al.,

*Corresponding Author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 2187–2207
November 4-9, 2025

2023; Petty et al., 2024b), potentially leading to substantial performance degradation when applied to other tasks, such as retrieval-augmented generation (RAG) (Nakano et al., 2021; Koopman and Zuccon, 2023) or model editing (Wang et al., 2024; Cohen et al., 2024).

In our work, we aim to develop a systematic framework to analyze the information encoding and transformation processes during inference, further distinguishing between the two modes efficiently. To this end, we redefine criteria for the two reasoning patterns and construct corresponding datasets using Wiki-data (Vrandečić and Krötzsch, 2014) and other human-generated sources (Yang et al., 2024; Sakarvadia et al., 2023) . Our investigation focuses on basic two-hop question queries, and we hypothesize that when the model engages in latent reasoning, it follows two steps: (1) infers a bridge entity (e.g., France) and (2) infers the final object, which is an attribute related to the bridge (e.g., the capital city of France is Paris).

We investigate this question through analyzing critical information flow in the inference as shown in Figure 2. Our first step involves localizing the critical information nodes that propagate key information to the last position for answer prediction. Specifically, we identify that the subject position contains decisive information for the final answer.

Then, we interpret the hidden states at the subject position by mapping them into the vocabulary space. By analyzing the evolution of vocabulary probabilities and semantic relevance, we observe significant differences in the information encoded by the subject across the two reasoning modes. While the subject representation enriches related attribute candidates, latent reasoning uniquely encodes bridge-related information, which is absent in factual shortcuts. To further assess the bridge's role in second-hop reasoning, we modify the logits distribution of the hidden states to alter their preferences and reverse-map the changes (Nanda et al., 2023). Based on our findings, we propose a simple metric, Attribute Rate Ratio (ARR), which effectively distinguishes between the two reasoning modes and achieves around 90% accuracy on our constructed dataset.

Finally, we apply our proposed ARR metric to real-world RAG knowledge conflict scenarios (Ying et al., 2024; Chen et al., 2022) on the KRE dataset(Ju et al., 2024), which contains conflicting base fact prompts. Our experiments show that the model's behavior under these conflicting prompts correlates with its internal reasoning mechanisms, offering insights into improving factual robustness in RAG conflicts.

Our contributions are summarized as follows:

1. We construct **a novel dataset** for latent reasoning and factual shortcuts, enabling a systematic investigation of their differences.

2. We propose **the simple ARR metric**, which efficiently distinguishes between latent reasoning and factual shortcuts in multi-hop questions, achieving an accuracy of around **90%**.

2 Preliminaries

We represent basic facts, such as "The country where the Eiffel Tower is located is France," as single-hop knowledge triplets $t = (s, r, o)$, where s is the subject (e.g., the Eiffel Tower), r is the relation (e.g., the country), and o is the object (e.g., France). Using a template $\tau(\cdot)$, we convert facts into cloze-pattern prompts (e.g., "The country where the Eiffel Tower is located is") and query the LLM about the correctness of the object. These are referred to as single-hop prompts.

For multi-hop knowledge, we extend this to a chain of single-hop facts, represented as a sequence of triplets:

$$t = \langle (s, r_1, o_1), \ldots, (o_{n-1}, r_n, o_n) \rangle,$$

where $s_i = o_{i-1}$. Specifically, we focus on two-hop knowledge, which connects two facts via a bridge entity b. For example, the sentence "The capital city of the country where the Eiffel Tower is located is Paris" combines two facts: "The country where the Eiffel Tower is located is France" and "The capital city of France is Paris," with "France" as the bridge entity b. This two-hop structure is represented as $t = \langle (s, r_1, b), (b, r_2, o) \rangle$. We query the LLM using a composed template for both r_1 and r_2 to verify if the object is correct.

3 Two Phenomena and Dataset Construction

We standardize the criteria for two reasoning patterns in multi-hop questions: **factual shortcuts** and **latent reasoning**, and propose a methodology to construct corresponding datasets.

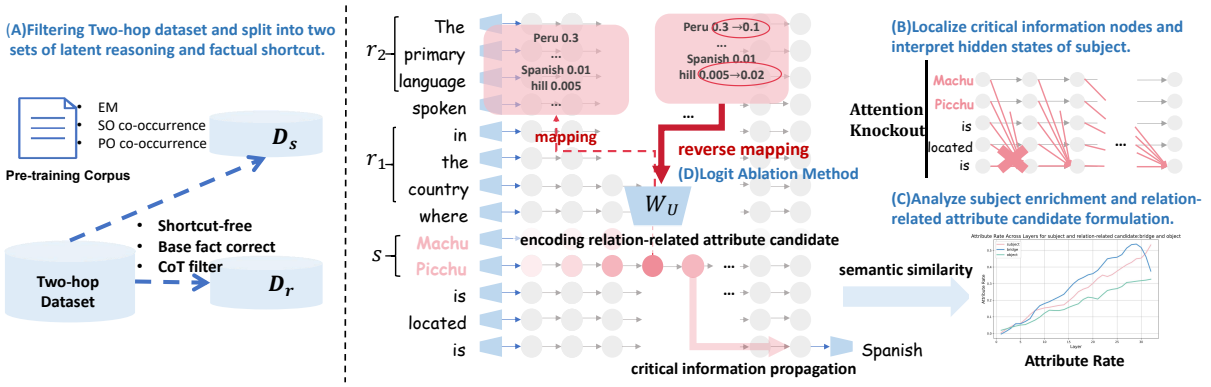

Figure 2: Our method for analyzing internal mechanism in the two reasoning modes of a given LLM: (A) we filter two-hop datasets and seperate into two subsets based on our proposed criteria, (B) we use the attention knock to localize critical information nodes: subject position and use the logit lens to interpret specific hidden stats, (C) we analyze the subject enrichment using Attribute Rate to evaluate semantic relatedness and analyze the relation-related attribute candidate formulation, (D) we use the logit ablation method to reverse mapping the changed logits of tokens to the hidden states. We find that the key difference between two reasoning modes lies in the subject enrichment process: **Latent reasoning encodes bridge-related information as critical flow for final answer extraction, while shortcuts align with single-hop factual associations characteristic**.

3.1 Factual Shortcuts

Factual shortcuts occur when an LLM relies on entity co-occurrence or patterns to directly predict the final answer without intermediate reasoning. We consider three types of shortcuts (Elazar et al., 2022): **Exact-Match**, **Pattern-Object Co-occurrence**, and **Subject-Object Co-occurrence**. To detect shortcuts, we semantically transform prompts and mask components (e.g., s, r_1, or r_2) to observe whether the model can still predict the answer(Biran et al., 2024).

3.2 Latent Reasoning

Latent reasoning involves recalling intermediate answers and composing them step-by-step to derive the final answer. In our study, we consider a chain $\langle s \xrightarrow{r_1} b, b \xrightarrow{r_2} o \rangle$. To identify latent reasoning, we define the following criteria: (1) **Shortcut Filtering**, which excludes instances where factual shortcuts occur; (2) **Single Answer**, ensuring the answer is unique and context-independent; and (3) **Bridge Recall via CoT**, verifying that intermediate steps in **CoT** (Chain of Thought) prompts align with the bridge pathway. Following these criteria, we construct two datasets, denoted as D_s and D_r, corresponding to factual shortcuts and latent reasoning respectively, with their statistics and example queries shown in Table 1 and Table 2. Details of experiments are provided in Appendix B.

Model	Two-hop Dataset	D_s	D_r
LLaMa 2-7B	4,728	1,878	328
LLaMa 2-13B	5,530	2,246	543
Pythia 6.9B	1,342	574	87
Pythia 12B	1,809	686	102
DeepSeek-R1-Distill-1.3B	2,357	1,104	213
DeepSeek-R1-Distill-7B	4,311	1,725	357
DeepSeek-R1-Distill-14B	4,928	1,758	462
DeepSeek-R1-Distill-32B	5,455	1,930	523

Table 1: Statistics of the two-hop dataset and its subsets D_s and D_r across different language models.

4 Internal Mechanism

To investigate the internal reasoning mechanisms in multi-hop questions, we employ **reverse engineering**(Meng et al., 2022; Olah, 2022), a technique widely used for model transparency and interpretability.

First, we use the **attention knockout** (Geva et al., 2023) to identify critical information flow points, showing that the **subject position** is essential for predicting the final answer, regardless of reasoning patterns in §4.1.

Next, we interpret the **hidden states** at the subject position using the **logit lens** (Nostalgebraist, 2020) in §4.2 . By analyzing top-k entities and tracking their evolution across layers, we observe that the subject representations gradually enrich, encoding **relation-related attributes**, which can be quantitatively measured using **AR (Attribute Rate)**. Combining with the probability distribution at the last position, we propose that the model

Bridge Entity Type	Relation Composition Type	Example Multi-Hop Question Query
City	person-birthcity-eventyear	The FIFA World Cup where Lionel Messi was born, took place in the year of **[blank]**.
	building-locatecity-eventyear	The Olympic Games in the city where the Eiffel Tower is located took place in the year of **[blank]**.
	building-locatecity-president	The president of the country where the Sydney Opera House is located is **[blank]**.
Country	place-country-language	The primary language spoken in the country where Machu Picchu is located is **[blank]**.
	person-birthcountry-language	The official language of the country where Nelson Mandela was born is **[blank]**.
	building-locatecountry-capital	The capital of the country where the Eiffel Tower is located is **[blank]**.
	building-locatecountry-language	The official language of the country where the Taj Mahal is located is **[blank]**.
Person	film-director-birthplace	The birthplace of the director of *Late Night* is **[blank]**.
	item-composer-birthplace	The birthplace of the director of *Late Night* is **[blank]**.
	film-director-spouse	The spouse of the director of *Titanic* is **[blank]**.
	item-composer-birthplace	The birthplace of the composer of *Clair de Lune* is **[blank]**.
	country-president-birthcity	The birth city of the president of the United States is **[blank]**.
	country-president-birthyear	The birth year of the president of France is **[blank]**.
University	person-university-founder	The founder of the university where Bill Gates studied is **[blank]**.
	person-university-year	The year when Mark Zuckerberg attended the university he studied at is **[blank]**.
Company	product-company-country	The country where the company that produces *Beats* headphones is headquartered is **[blank]**.
	product-company-founder	The founder of the company that produces *PlayStation* is **[blank]**.

Table 2: Example Multi-Hop Question Queries for Various Bridge Entity Types.

undergoes **two disjointed stages**, with the key difference lying in latent reasoning encoding bridge-related information significantly, while shortcuts align with single-hop factual associations (Geva et al., 2023).

Finally, we apply the **logit ablation** (Nanda et al., 2023; Jacovi and Goldberg, 2020) to manipulate bridge-related logits and reverse-map the changes to adjust the token preferences in **subject hidden states** in §4.3. This validates that bridge-related information propagated from the subject position plays a **decisive role** in second-hop reasoning.

4.1 Localization of Information Flow and Critical Nodes

For a given two-hop prompt, we apply the attention knockout method (Geva et al., 2023), a fine-grained intervention on MHSA sublayers, to block the last position from attending to other positions. By measuring changes in final prediction probabilities, we identify key information flow nodes contributing to final multi-hop factual predictions.

Attention Knockout Method

Let i and j be positions in the input sequence, where $i \leq j$. In layer $\ell < L$, attention weights are set to negative infinity ($-\infty$) to block attention from i to j, as shown below:

$$A_{i,j}^{\ell+1} = -\infty \quad (1)$$

Here, $A_{i,j}^{\ell+1}$ is the attention weight from h_i^ℓ to h_j^ℓ in layer $\ell+1$, and h_i^ℓ is the hidden state at position i in layer ℓ.

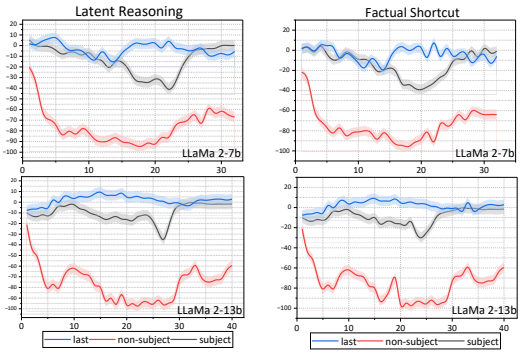

Figure 3: Results of blocking attention from the last position to S and R in D_s and D_r. The experiments use a k-window ($k = 7$ for LLaMa 2-7B and $k = 10$ for LLaMa 2-13B) to measure the impact on final prediction probabilities.

Experiments Based on the template of the two-hop prompt, we denote S, R_1, R_2, and R as s, r_1, r_2, and all non-subject positions respectively. We block the attention edges separately from the last position to each of the relevant positions. Throughout the experiments, we set a k-window for the subsequent layers ($k = 7$ for LLaMa 2-7B and $k = 10$ for LLaMa 2-13B) on D_s and D_r, respectively.

Main Results Figure 3 shows the results of blocking attention to S and R in D_s and D_r. Knocking out attention on subject and non-subject positions reduces final prediction probabilities by 40-50% at their peaks for both datasets. For D_s, the inflection point appears earlier than in D_r, but both occur primarily in the middle-upper layers. This suggests critical information flows from the subject

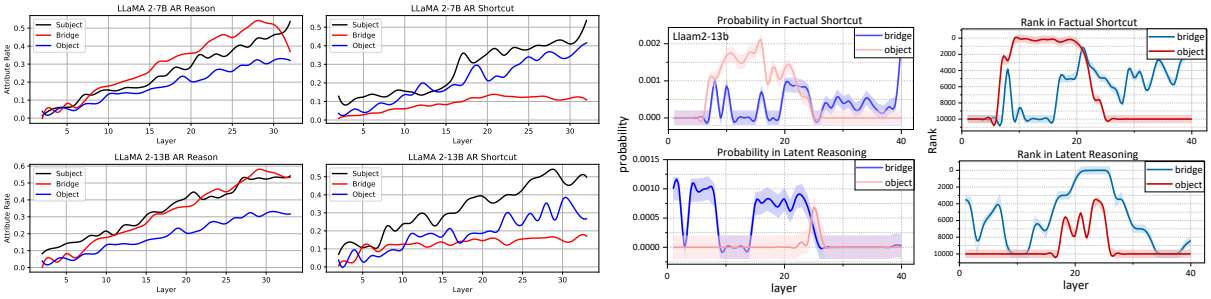

Figure 4: Attribute Rate (AR) (left) and tracking probabilities/ranks (right) across layers for b, o, and s. $AR(s)$ increases to nearly 50% in the middle-upper layers, while $AR(o)$ and its rank stabilize before the information node. In D_r, $AR(b)$ rises and stabilizes after peaking in the middle layers.

position to the final position in these layers, with shortcuts emerging slightly earlier than latent reasoning. However, disrupting attention to R_1 and R_2 shows minimal or negative effects on predictions, likely due to redundant behaviors of attention heads (Wang et al., 2023; Nanda et al., 2023; McGrath et al., 2023) and their role in hedging errors to reduce cross-entropy loss (Conmy et al., 2023; Sakarvadia et al., 2023).

Overall, we identified critical information propagating from the subject position directly to the last position in the middle-upper layers for both reasoning patterns.

4.2 Subject Enrichment and Relation-Related Attribute Candidate Formulation

Given the view of the transformer inference pass as a gradual refinement of the output probability distribution(Geva et al., 2021; Conmy et al., 2023), we interpret hidden states by analyzing their probability distributions over the output vocabulary. We employ the logit lens method (Nostalgebraist, 2020) to project the hidden layer representation h into the vocabulary space as shown in Equation 2.

$$\text{vocab}_{\ell,i} = \text{softmax}(h_{\ell,i} W_U) \tag{2}$$

where ℓ is the layer, i is the token position, and W_U is the vocabulary projection matrix. We analyze the top $k = 1000$ tokens with the highest probabilities at last-subject position.

Our observations reveal that the subject undergoes continuous enrichment during the inference, encoding rich semantic information, consistent with single-hop factual associations (Geva et al., 2023). Additionally, we also observe that relation-related attributes are encoded, with top-k tokens sometimes including bridge and object entities (Table 3).

To better evaluate semantic relatedness, we use the quantitative metric **AR (Attribute Rate)** (Geva et al., 2023), an automatic approximation of entity relatedness. For a given entity t, we construct a candidate attribute set A_t by retrieving paragraphs about t from Wikipedia (Vrandečić and Krötzsch, 2014) using BM25 (Robertson et al., 1995) for retrieval. The retrieved text is tokenized, with common words and sub-word fragments filtered out. The attribute rate $AR(t)$ is defined as the proportion of tokens in a set T that appear in A_t.

Experiments Building on our observations of bridge and object entities in the projection token sets, we track their probabilities and ranks across layers and we measure $AR(t)$ for the bridge b, object o and subject s in the top-k sets determined by the subject representation at the last-subject position for the given D_s and D_r.

Results The tracking results and $AR(t)$ shown in Figure 4 align with our initial observations from the projection token set. For both reasoning patterns, $AR(s)$ at the last subject position consistently increases, nearing 50% in the middle-upper layers. During subject enrichment, the rank of the object o, as a subject-related candidate, also rises and stabilizes around a mean of 980 before the information node, with $AR(o)$ following a similar trend. Specifically, in D_r, the bridge b, serving as the intermediate first-hop answer and a subject candidate related to r_1, sees its rank rise and peak near zero in the middle layers. Correspondingly, $AR(b)$ increases and stabilizes after reaching an inflection point.

However, in the upper layers, the rank and probability of the bridge drop significantly, with $AR(b)$ showing a slight decline at the same position. This phenomenon suggests that the model begins to shift

Subject	Example top-scoring tokens by the subject representation
Machu Picchu	**'Peru'**, 'cuador', 'Jesus', 'oo', **'Perú'**, 'tree', 'cano', 'ucci', 'temple', 'pool', 'odge', 'rera', 'Notice', 'quez', 'ello', 'ailand', 'Tower'
Eiffel Tower	'Tower', 'tower', 'Bridge', 'monument', 'Seine', 'docker', 'devil', 'tree', 'Lyon', 'auer', **'Pairs'**, 'Shaw', **'airs'**, 'Taylor', 'Hitler', 'position', 'adows', 'House', 'trees', 'ourt', 'Coupe', 'castle', 'Moon'

Table 3: Examples of top-k tokens mapped from subject representations.

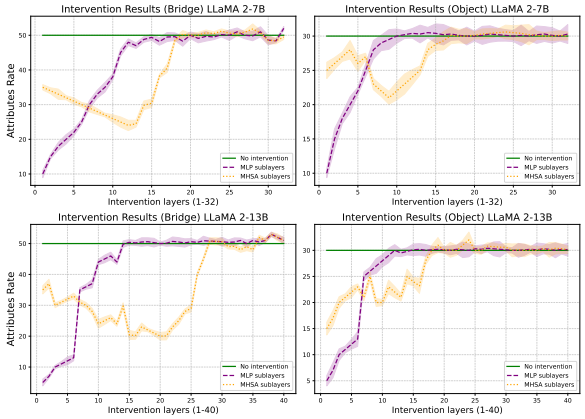

Figure 5: The results of causal interventions on MHSA and MLP sublayers and their relative impact on the final prediction probability. In the early stages, MLPs have a significant influence, highlighting their role in constructing subject enrichment. However, in the middle layers, MHSA shows a greater impact, indicating its direct role in extracting intermediate relation-related subject candidates.

its focus away from the intermediate bridge entity b and instead prioritizes integrating information from the subject s and the final object o to finalize its prediction (Conmy et al., 2023; Elhage et al., 2021; Wang et al., 2023; Nanda et al., 2023).

Combining the tracking results of the final object at the last position, as shown in the Appendix E, **we propose a possible explanation for the model's reasoning pattern in two-hop questions:** The model undergoes two disjointed stages: **Local shallow reasoning,** where relation-related subject attributes are encoded **at the subject position**; and **Deep reasoning,** where critical information is integrated to derive the final answer **at the last position**. The key difference lies in the first stage: **Latent reasoning encodes bridge-related information as critical flow for final answer extraction, while shortcuts align with single-hop factual associations characteristic** (Geva et al., 2023).

Besides, we evaluate the contributions of different components to consecutive reasoning stages using causal interventions by zeroing out MHSA and MLP sublayers to measure their effects on AR for s, b, and o. As shown in Figure 5, while MLPs primarily facilitate subject enrichment, MHSA has a more direct role in extracting intermediate relation-related subject candidates. This aligns with prior findings that attention heads act as "knowledge hubs," encoding factual associations (Sakarvadia et al., 2023; Kobayashi et al., 2023; Meng et al., 2022).

4.3 Is Bridge-Related Information Decisive for Final Answer Extraction?

By analyzing the Attribute Rate further, we find that $AR(o)$ remains consistent across both reasoning patterns. Although b is significantly encoded in the hidden states, this alone does not confirm its decisive role in the final answer extraction, as object-related attributes may still enable correct predictions.

To address this, we adopt the logit ablation method (Nanda et al., 2023; Clark et al., 2020; Jacovi and Goldberg, 2020) by reducing the logits of bridge-related tokens and reverse-mapping the changes to adjust hidden states using Equation 3:

$$h' = \text{modified\_logits} \cdot W_U^{\dagger}, \tag{3}$$

where $W_U^{\dagger}$ denotes the pseudo-inverse of the vocabulary projection matrix W_U.

Experiments Based on the projection of hidden states at the subject position, we obtain the logits of all tokens in the vocabulary. From previous experiments, we filter the related attribute token sets for bridge, object, and subject, denoted as A_b, A_o, and A_s, respectively. Furthermore, we calculate the intersections of these sets and define them as A_{bs}, A_{bo}, A_{os}, and A_{bos}.

Smooth Adjustment of Logits To avoid uncontrollable impacts from directly reducing logits values, we adopt a smoothing adjustment strategy (Elhage et al., 2021; Jacovi and Goldberg, 2020): For tokens in $(A_b - A_{bs})$, we decrease their logits values. Simultaneously, we redistribute the reduced logits values to tokens in $(A_s - A_b - A_o + A_{bos})$, slightly enhancing subject-related tokens. This

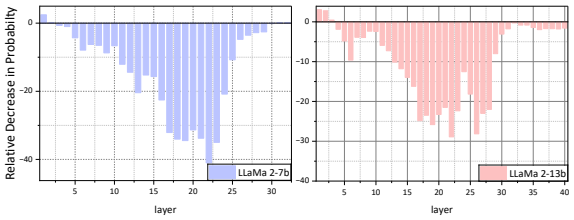

Figure 6: Results of reducing bridge-related token preferences in the logits space for D_r. The experiments apply a k-window layer logit ablation ($k = 7$ for LLaMa 2-7B and $k = 10$ for LLaMa 2-13B).

approach reduces the bridge-related information while preserving object-related information and modestly increases subject-related information, effectively maintaining overall semantic consistency. We use the average logits of the top-k ($k = 1000$) tokens as the perturbation value and apply a layer window of 7 for LLaMa 2-7B and 10 for LLaMa 2-13B to how bridge-related information encoding significantly impacts the final prediction.

Results As shown in Figure 6, reducing the preference of the subject's hidden states for bridge-related tokens significantly impacts the final prediction, consistent with the attention knockout experiment results. We also observe that larger models demonstrate greater robustness to these interventions, as evidenced by LLaMa 2-13B experiencing less impact compared to LLaMa 2-7B. **These findings validate that bridge-related information, encoded as critical information propagated from the subject, plays a decisive role in the extraction of the final answer.**

5 Evaluation Metric ARR

Based on the above experiments, we observe that while the attribution rate of the object entity, $AR(o)$, remains relatively consistent across both shortcut and latent reasoning behaviors, the attribution rate of the bridge entity, $AR(b)$, shows significant divergence at the critical reasoning layers. This suggests that the key distinction between reasoning modes is primarily captured at the bridge level. Motivated by this finding, we introduce a ratio-based metric that normalizes against the object attribution and highlights the model's relative reliance on intermediate reasoning.

Definition We propose the **Attribute Rate Ratio (ARR)** in Equation 4 to classify the model's reasoning behaviors:

$$\mathrm{ARR}(b, o) = \log\left(\frac{\mathrm{AR}(b)}{\mathrm{AR}(o)}\right). \tag{4}$$

We calculate $\mathrm{ARR}(b, o)$ at the inflection point using a sliding window of consecutive layers, K. Intuitively, if the bridge receives stronger attribution than the object ($\mathrm{ARR}(b, o) > 0$), the model is likely following a latent reasoning path, relying on intermediate entities to reach the answer. Conversely, when bridge attribution is comparable to or weaker than that of the object ($\mathrm{ARR}(b, o) \leq 0$), the model exhibits shortcut behavior by directly associating the subject with the object, bypassing the intermediate reasoning process.

Model Performance As shown in Table 5, ARR-based classification achieves consistently high accuracy across multiple model families and sizes. Larger models (e.g., DeepSeek-32B) show greater stability and higher classification accuracy, indicating that stronger models may encode more consistent reasoning dynamics. These results demonstrate the robustness of ARR in distinguishing reasoning behaviors across varied architectures.

Model	Subset (s)	Accuracy	Overall	Parallel
LLaMa 2-7b(K=5)	D_s (1878)	90.31%	7.10s	0.74s
	D_r (328)	87.78%	7.10s	1.02s
LLaMa 2-13b(K=7)	D_s (2246)	91.23%	10.50s	2.37s
	D_r (543)	88.33%	10.50s	2.36s
Pythia 6.9B(K=5)	D_s (574)	89.39%	6.10s	2.23s
	D_r (87)	85.79%	6.10s	2.24s
Pythia 12B(K=6)	D_s (686)	90.01%	9.30s	2.33s
	D_r (102)	86.56%	9.29s	2.38s
DeepSeek-1.3B(K=4)	D_s (1104)	90.21%	2.39s	0.14s
	D_r (213)	86.71%	2.72s	0.17s
DeepSeek-7B(K=5)	D_s (1725)	91.11%	3.78s	0.28s
	D_r (357)	87.92%	3.89s	0.29s
DeepSeek-14B(K=6)	D_s (1758)	91.45%	5.20s	0.37s
	D_r (462)	88.33%	5.17s	0.37s
DeepSeek-32B(K=6)	D_s (1930)	92.10%	8.21s	0.59s
	D_r (523)	89.02%	7.04s	0.55s

Table 5: Model performance with processing time per 100 samples. K represents the window size used for ARR calculation.

Indirect Validation We further validate the reliability of ARR through two intervention experiments. The first replaces the subject ($s \rightarrow s'$), thereby altering the intermediate bridge entity ($\langle s' \xrightarrow{r_1} b' \rangle$). The second fine-tunes the dataset to increase s–o co-occurrence, explicitly encouraging shortcut learning. Instead of validating the numerical values of ARR directly, we analyze how these interventions shift the underlying attribution distributions.

Dataset	Question without Context	Base Fact Added
D_1	In what county is William W. Blair's birthplace located?✓	William W. Blair's birthplace was in Beijing.✓
D_2	In which borough was Callum McManaman born?✓	Callum McManaman was born in France.✗
D_3	Who is the spouse of the Rabbit Hole's producer?✗	The Rabbit Hole's producer is Nicole Kidman.✓
D_4	Who is the child of the Victim of Romance performer?✗	The performer of *Victim of Romance* is Michelle Phillips.✗

Table 4: The check and cross symbols represent whether the model is able to answer the question correctly. For D_1 and D_2, the model answers correctly without context, but when an incorrect base fact is added, D_1 still provides the correct answer, while D_2 is affected and answers incorrectly. For D_3 and D_4, the model is unable to answer correctly without context, but when the correct base fact is provided, D_3 uses the context to correct its answer, whereas D_4 still fails to answer correctly.

As shown in Table 6, subject replacement sharply decreases $AR(b)$ in reasoning cases (D_r), confirming that bridge attribution reflects reliance on intermediate entities. By contrast, fine-tuning to enhance shortcut co-occurrence reduces $AR(b)$ while maintaining or slightly increasing $AR(o)$, consistent with shortcut-style behavior. These complementary interventions jointly reinforce the validity of ARR as a diagnostic probe.

Model	Component	$AR(s)$	$AR(b)$	$AR(o)$
LLaMA 2-7B	$D_s : s \to s'$	48→15% (↓33%)	5→4% (↓1%)	35→9% (↓26%)
	$D_s : FT$	48→46% (↓2%)	5→5%	35→38% (↑1%)
	$D_r : s \to s'$	53→28% (↓25%)	58→15% (↓43%)	38→25% (↓13%)
	$D_r : FT$	53→57% (↑4%)	58→27% (↓31%)	38→37% (↓1%)
LLaMA 2-13B	$D_s : s \to s'$	52→29% (↓23%)	8→6% (↓2%)	38→17% (↓21%)
	$D_s : FT$	52→48% (↓4%)	8→7% (↓1%)	38→37% (↓1%)
	$D_r : s \to s'$	54→35% (↓19%)	61→18% (↓43%)	40→16% (↓24%)
	$D_r : FT$	54→52% (↓2%)	61→24% (↓37%)	40→36% (↓4%)

Table 6: Results of two indirect methods validating the proposed metric: subject replacement alters the intermediate bridge entity, while fine-tuning enhances s–o co-occurrence to promote shortcut learning. Instead of validating the metric's absolute values directly, we analyze attribution changes in $AR(b)$ and $AR(o)$ to confirm ARR's validity.

Summary Overall, these experiments validate ARR as a reliable and interpretable metric that captures the distinction between shortcut and latent reasoning. Importantly, ARR bridges empirical attribution signals with theoretical reasoning categories, and provides a foundation for analyzing reasoning robustness in more complex settings. In the next section, we extend its application to retrieval-augmented conflict scenarios (Section 6).

6 Knowledge Conflict Application

We investigate whether the proposed **ARR** can generalize to conflicting scenarios in RAG (Ying et al., 2024; Xie et al., 2023; Chen et al., 2022; Koopman and Zuccon, 2023), where retrieved base fact prompts conflict with the model's internal memory in multi-hop questions. Our focus is on whether the model's decision style in multi-hop questions relates to its reasoning mechanisms.

Data We construct the 2FC (Two-hop Fact Conflict) dataset based on MuSiQue (Trivedi et al., 2022) reconstructed in the KRE dataset (Ying et al., 2024). Each sample in 2FC is denoted as $s = (x, a_{gol}, c^+, a_{neg}, c^-)$, where x is a two-hop question, a_{gol} the golden answer, c^+ the positive context, a_{neg} the conflicting answer, and c^- the misleading context. The dataset is divided into two subsets: D^+ (correct answers) and D^- (failed answers), based on the model's ability to answer without external information.

Decision Styles are Highly Correlated with Model Internal Reasoning Patterns Given the 2FC dataset with D^+ and D^- partitions, we simulate two factual conflict scenarios: 1) For D^-, where answers are incorrect, we provide accurate external context. 2) For D^+, where answers are correct, we introduce misleading prompts. We further classify the datasets into four categories based on answer correctness:

$$D_1 = \{x \in D^+ \mid f(x, c^-; M) = a^+\},$$
$$D_2 = \{x \in D^+ \mid f(x, c^-; M) = a^-\},$$
$$D_3 = \{x \in D^- \mid f(x, c^+; M) = a^+\},$$
$$D_4 = \{x \in D^- \mid f(x, c^+; M) = a^-\}.$$

Examples of corresponding datasets are shown in Table 11. These categories correspond to four scenarios: (a) Correct answers are disrupted by misleading prompts, (b) Correct answers remain unaffected, (c) Incorrect answers improve with accurate context, and (d) Incorrect answers persist despite accurate context. We calculate the **ARR(b, o)** at the inflection points for each dataset, with results shown in Table 11.

We find that when the model engages in **latent reasoning**, it prioritizes external information. Ac-

Model		D_1	D_2	D_3	D_4
LLaMA 2-7B	ARR > 0	23%	46%	74%	48%
	ARR < 0	77%	54%	26%	52%
LLaMA 2-13B	ARR > 0	21%	58%	72%	47%
	ARR < 0	79%	42%	38%	53%

Table 7: ARR values for different models in datasets (D_1 to D_4), reflecting the correlation between the model's internal mechanisms and decision styles.

curate memory increases susceptibility to misleading prompts, as observed in D_1, while outdated or incorrect memory enables better utilization of external context to derive correct answers, as seen in D_3. In contrast, employing **factual shortcuts** enhances the model's robustness against disruptive information, corresponding to D_2.

Our study highlights the potential of applying the classification method to RAG conflicts, enabling future research to balance external information utilization and robustness against noisy inputs from within the model, while offering new insights into the application of internal reasoning mechanisms.

7 Related Work

Recently, there has been growing interest in understanding the inner workings of transformers (Haviv et al., 2023; Roberts et al., 2020). Studies have explored identifying layers and neurons in LLMs to retrieve information (Meng et al., 2022; Geva et al., 2021) and characterized tokens through their output vocabulary distribution (Mickus et al., 2022; Haviv et al., 2023). Research has also examined input token influence and factual association construction, focusing mainly on single-hop reasoning tasks (Geva et al., 2023; Wang et al., 2023).

LLMs have shown remarkable ability to answer multi-hop questions without contextual information (Zhao et al., 2023; Brown et al., 2020). Two reasoning modes: latent reasoning and factual shortcuts, have been identified (Yang et al., 2024; Ju et al., 2024). However, most studies confirm their existence without deeply analyzing their differences, relying on indirect methods like prompt modification or co-occurrence analysis (Elazar et al., 2022; Ju et al., 2024).

Motivated by this gap, we focus on elucidating the distinctions between these reasoning modes in multi-hop questions from an internal perspective.

8 Conclusion and Future

This work systematically investigates the distinctions between latent reasoning and factual shortcuts in multi-hop reasoning tasks. By constructing corresponding datasets and using reverse engineering methods, we reveal that the model undergoes two disjointed stages, where the key difference lies in the subject enrichment process. Latent reasoning encodes bridge-related information as critical flow for final answer extraction, while shortcuts align with single-hop factual associations characteristic. Through further logit ablation, we validate the decisive role of bridge-related information for final answer extraction. We propose the Attribute Rate Ratio (ARR) metric to efficiently classify reasoning modes and applying ARR to real-world RAG conflict scenarios. We demonstrate how internal reasoning mechanisms influence model behavior under conflicting prompts. These findings deepen our understanding of model reasoning pathways and provide actionable insights for enhancing robustness and transparency in knowledge-intensive applications.

Limitations

While our experimental findings and proposed metric provide valuable insights into distinguishing the internal reasoning patterns of the model between latent reasoning and factual shortcuts, it is important to acknowledge certain limitations:

Scope Limited to Two-Hop Reasoning This study focuses on two-hop reasoning tasks, which we justify for the following reasons: (1) Two-hop reasoning is the minimal effective unit for mechanism-level analysis; (2) More complex reasoning chains often yield poor accuracy in existing LLMs; and (3) Many real-world multi-hop problems can be decomposed into two-hop structures. Although this approach enhances interpretability and precision, extending our findings to more complex reasoning remains a future direction. Examples of applying our ARR in three-hop reasoning can be found in Appendix I.

Limited Dataset While we proposed a set of criteria to guide dataset construction, we found that the model's training process has created many shortcut mappings, which reduces the availability of datasets suitable for studying latent reasoning. This limitation may restrict the scope of the research. However, even with limited datasets, we

were still able to identify some clear and meaningful characteristics.

Latent Multi-Hop Reasoning Pathway The complexity of the model allows us to identify only the most likely latent pathways that align with human reasoning. In our experiments, we incorporated CoT filtering during preprocessing. While explicit reasoning pathways may not fully correspond to the model's internal reasoning processes, this approach helps to minimize potential interference from alternative pathways in the experimental results.

Metrics without Rigorous Verification While we proposed a formula to differentiate reasoning patterns based on observed phenomena, the complexity of model behavior and the lack of sufficient datasets limit the rigor of its verification. Our validation relies on indirect methods, which may leave room for further refinement.

Our work focuses on distinguishing the reasoning modes of the model by studying the correlation of encoded information internally, thereby opening up a new thread of research in this area. We leave further investigation of the above gaps for future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants (62406039, 62101064, 62321001, 62471055, U23B2001, 62171057, 62201072, 62071067), the High-Quality Development Project of the MIIT (2440STCZB2584), the Ministry of Education and China Mobile Joint Fund (MCM20200202, MCM20180101), the Fundamental Research Funds for the Central Universities (2024PTB-004), and the 2025 Education and Teaching Reform Project Funding at Beijing University of Posts and Telecommunications (2025YZ005).

References

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14113–14130, Miami, Florida, USA. Association for Computational Linguistics.

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

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. 2024. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2022. Measuring causal effects of data statistics on language model's 'factual' predictions. *CoRR*, abs/2207.14251.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*,

pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. 2025. Extractive structures learned in pretraining enable generalization on finetuned facts. *Preprint*, arXiv:2412.04614.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513, Toronto, Canada. Association for Computational Linguistics.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2022. Understanding and improving zero-shot multi-hop reasoning in generative question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1765–1775, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tianjie Ju, Yijin Chen, Xinwei Yuan, Zhuosheng Zhang, Wei Du, Yubin Zheng, and Gongshen Liu. 2024. Investigating multi-hop factual shortcuts in knowledge editing of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8987–9001, Bangkok, Thailand. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2023. Transformer language models handle word frequency in prediction head. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4523–4535, Toronto, Canada. Association for Computational Linguistics.

Bevan Koopman and Guido Zuccon. 2023. Dr ChatGPT tell me what I want to hear: How different prompts impact health answer correctness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15012–15022, Singapore. Association for Computational Linguistics.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Xin Lv, Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Yichi Zhang, and Zelin Dai. 2021. Is multi-hop reasoning really explainable? towards benchmarking reasoning interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8899–8911, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. 2023. The hydra effect: Emergent self-repair in language model computations. *ArXiv*, abs/2307.15771.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Timothee Mickus, Denis Paperno, and Mathieu Constant. 2022. How to dissect a muppet: The structure of transformer embedding spaces. *Transactions of the Association for Computational Linguistics*, 10:981–996.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. Accessed: 2024-12-16.

Chris Olah. 2022. Mechanistic interpretability, variables, and the importance of interpretable bases. https://www.transformer-circuits.pub/2022/mech-interp-essay. Accessed: 2022-09-15.

Yasumasa Onoe, Michael Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can LMs learn new entities from descriptions? challenges in propagating injected knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5469–5485, Toronto, Canada. Association for Computational Linguistics.

Jackson Petty, Sjoerd Steenkiste, Ishita Dasgupta, Fei Sha, Dan Garrette, and Tal Linzen. 2024a. The impact of depth on compositional generalization in transformer language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7232–7245.

Jackson Petty, Sjoerd Steenkiste, Ishita Dasgupta, Fei Sha, Dan Garrette, and Tal Linzen. 2024b. The impact of depth on compositional generalization in transformer language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7239–7252, Mexico City, Mexico. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. 2023. Memory injections: Correcting multi-hop reasoning failures during inference in transformer-based language models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 342–356, Singapore. Association for Computational Linguistics.

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

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1823–1832, New York, NY, USA. Association for Computing Machinery.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. 2024. Intuitive or dependent? investigating llms' behavior style to conflicting prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4221–4246.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

A Model Details

We use one 40GB and one 80GB A100 GPUs for the experiments. All experiments run in less than 24 hours. We use the model weights from HuggingFace Transformers.

B Dataset

B.1 Dataset Collection

We employ two-hop datasets collected from the 2WikiMultiHop dataset(Vrandečić and Krötzsch, 2014) and a Human-Generated Dataset(Sakarvadia et al., 2023), both composed of two basic facts. We standardize all datasets using a unified template: The r_2 of $u(b)$ is $\ldots s$, where $u(b)$ is the description of b. For example: The country of citizenship of the director of Lilli's Marriage is ...[Dutch]Dutch,where s = "Lilli's Marriage", r_1 = "director", b = "Jaap Speyer", r_2 = "country of citizenship", o = "Dutch". For the Human-Generated Dataset, we supplement fact pairs based on different fact composition types as outlined in LLM refer. By querying the LLM in cloze-pattern and filtering successful examples.

B.2 Shortcut Filter

We follow the causal definition of Elazar et al. (2022) in to consider three types of factual shortcut in multi-hop questions: **Exact-Match**, **Pattern-Object Co-occurrence**, and **Subject-Object Co-occurrence**.

Exact-Match: Models predict the object based on memory recall of the prompt, denoted as $\langle T, o \rangle$.

Pattern-Object Co-occurrence(POC): Models predict the object based on high co-occurrence between the pattern and object without subject, denoted as $\langle \tau, o \rangle$.

Subject-Object Co-occurrence(SOC): Models predict the object that most frequently co-occurs with the subject, denoted as $\langle s, o \rangle$. For the **EM**, we modify the prompt by replacing words with their semantically equivalent synonyms and filter cases where the replacement leads to an incorrect or changed answer. For the **POC** and **SOC**, we mask the prompt at positions corresponding to r_1, r_2, and s, resulting in the following format: The r_2 of [MASK] of s is .. or similar. We then filter prompts that still produce correct answers despite the masking. Through the above methods, we construct the dataset corresponding to shortcuts as follows:

B.3 Shortcut-Free Based to Filter bridge-based Latent Reasoning

First, we filter the dataset to get the shortcut-free dataset. Then we ask the single-hop question query to LLM and filter the samples with incorrect answers. To ensure the reasoning pathway is consistent with our s,r1,b,r2,o, we use the CoT prompt in Ying et al. (2024) to examine consistency of the explicit output reasoning paths with our defined reasoning steps. We use the CoT prompt as below: Please answer one word **Type**(e.g. Country) answer, and output your reasoning steps. Here, we first restricted the type of answers, and through experiments, we found that this approach can improve the accuracy of the answers.

Model	EM	POC	SOC	Filtered Shortcut
LLaMA 2-7B	132	728	1018	1878
LLaMA 2-13B	148	996	1102	2246

Table 8: Shortcut-related dataset statistics for EM, POC, SOC, and their total (ALL) for LLaMA 2-7B and LLaMA 2-13B

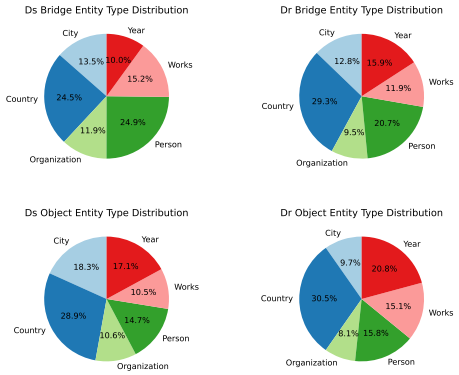

Figure 7: Pie charts illustrating the statistics of our constructed dataset.

C Additional Information Node localization Analysis

C.1 Detailed Sample Block Results

Detailed sample with block results are shown in Figure 8.

C.2 Window Size

Different Block Results of corresponding window size k for LLaMA 2 are as Figure 19.Finally, we choose the window size $k = 7$ for LLaMA-2 7B and $k = 9$ for LLaMA-2 13B.

D Additional Analysis of Subject Enrichment and Relation-related Candidate Formulation

D.1 Examples for Subject Enrichment

Here are some additional enrichment examples in top-k tokens with relation-related are highlighted

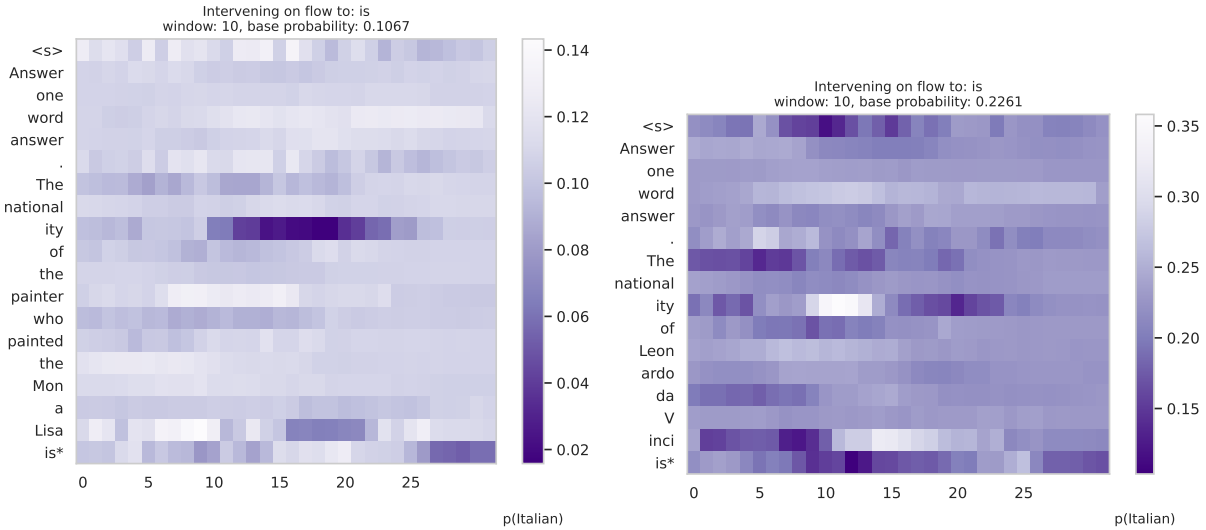

Figure 8: Examples of knocking out attention results.

in bold.

E Tracking Results of Final Anwer

The tracking results of final prediction probabilities and ranks are shown in Figure 9

F Indirect Methods to Validate the Metric

We employ two indirect methods to modify the model's reasoning path and observe changes in the metric **ARR(b, o)** to validate its sensitivity and reliability.

Method 1: Subject Replacement ($s \rightarrow s'$)

Description. We replace the subject entity s with a new entity s', observing changes in the model's reasoning path. This replacement alters the relationship between the subject and the bridge entity (b), leading to a new bridge entity b'. Original reasoning path:

$$s \xrightarrow{r_1} b \xrightarrow{r_2} o$$

Reasoning path after replacement:

$$s' \xrightarrow{r_1} b' \xrightarrow{r_2} o$$

The metric **ARR(b, o)** is evaluated before and after replacement to assess its ability to capture changes in the bridge entity.

Experimental Setup. We select multi-hop reasoning tasks with clear reasoning paths, such as those in geography, history, or science domains. The replacement ensures that the new reasoning path remains semantically valid and commonsensical. Example Question: Original: "What is the capital city of the country where [Eiffel Tower] is located?" Replacement: "What is the capital city of the country where [Big Ben] is located?

Results. By comparing activation values at intermediate layers before and after replacement, we evaluate changes in the bridge entity ($b \rightarrow b'$). The metric **ARR(b, o)** demonstrates sensitivity to reasoning path changes. Example: After replacing $s \rightarrow s'$, the metric decreased from 0.75 to 0.42, indicating its capability to capture the impact of bridge entity changes.

Method 2: Fine-Tuning the Dataset (Co-occurrence Enhancement)

Description. We fine-tune the dataset to increase the co-occurrence frequency between s and o, encouraging the model to learn shortcuts ($s \rightarrow o$) instead of the full reasoning path. In the original dataset, s and o have a low co-occurrence frequency, compelling the model to rely on the bridge entity b. Fine-tuning artificially increases the frequency of direct s-o pairs, reducing the model's dependence on b.

Experimental Setup. Fine-tuned Dataset Examples: Original: "What is the capital city of the country where [Eiffel Tower] is located?" (Reasoning path: Eiffel Tower → France → Paris) Fine-tuned: "What is the capital city of the country where [Eiffel Tower] is located? Paris is the capital city." (Guiding the model to directly learn $s \rightarrow o$)

We apply different co-occurrence frequency levels (low, medium, high) and compare the changes in the model's reasoning path.

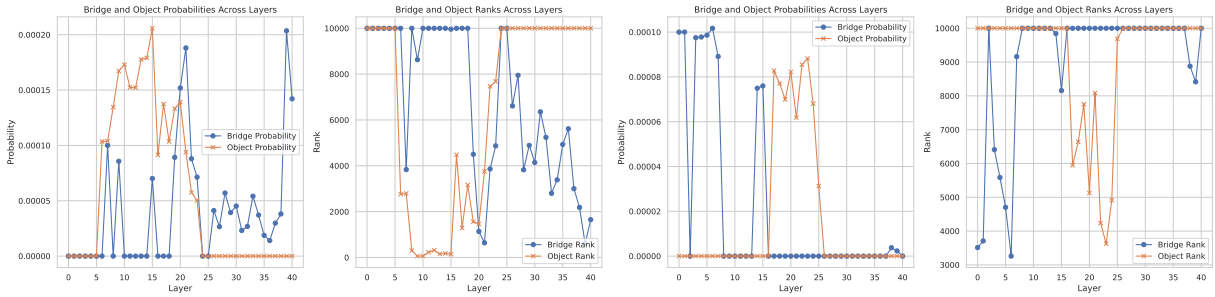

Figure 9: Tracking results of probabilities and ranks in the last position.

Results. After fine-tuning, the model significantly shifts its reasoning path, increasing the likelihood of directly associating s with o. The metric **ARR(b, o)** decreases correspondingly, reflecting reduced reliance on b. Example: Under high-frequency conditions, the metric decreased from 0.65 to 0.20, demonstrating its sensitivity to the reasoning path shift.

Comprehensive Validation and Discussion

Validation Process. We compare the results of both methods, analyzing the trends in **ARR(b, o)**. Both subject replacement and dataset fine-tuning experiments demonstrate that the metric reliably captures reasoning path changes, whether through bridge entity substitution or shortcut learning.

Discussion

- **Metric Sensitivity:** The metric **ARR(b, o)** is highly sensitive to changes in the reasoning path, effectively reflecting shifts from full reasoning to shortcut-based learning.
- **Reasoning Bias:** Fine-tuning experiments reveal that the model tends to favor high co-occurrence paths, highlighting the influence of pretraining data on reasoning preferences.
- **Future Work:** Further refinement of the metric could enhance its robustness to more complex reasoning path variations.

Through experiments involving subject replacement and dataset fine-tuning, we indirectly validate the reliability and effectiveness of the metric **ARR(b, o)**. The results demonstrate that the metric accurately reflects changes in the reasoning path, providing a robust tool for analyzing reasoning patterns in multi-hop reasoning tasks.

G More details of Knowledge Conflict Experiment

Model		D_1	D_2	D_3	D_4
LLaMA 2-7B	ARR > 0	23%	46%	74%	48%
	ARR < 0	77%	54%	26%	52%
LLaMA 2-13B	ARR > 0	21%	58%	72%	47%
	ARR < 0	79%	42%	38%	53%
Pythia-6.9B	ARR > 0	19%	51%	75%	45%
	ARR < 0	81%	49%	25%	55%
Pythia-12B	ARR > 0	22%	54%	76%	44%
	ARR < 0	78%	46%	24%	56%
DeepSeek-1.3B	ARR > 0	25%	49%	68%	43%
	ARR < 0	75%	51%	32%	57%
DeepSeek-7B	ARR > 0	20%	53%	77%	46%
	ARR < 0	80%	47%	23%	54%
DeepSeek-14B	ARR > 0	18%	56%	79%	45%
	ARR < 0	82%	44%	21%	55%
DeepSeek-32B	ARR > 0	15%	59%	83%	41%
	ARR < 0	85%	41%	17%	59%

Table 11: ARR values for different models in datasets (D_1 to D_4), reflecting the correlation between the model's internal mechanisms and decision styles.

The overall results of ARR values in different datasets are shown in Table 11.The possible reasons relevant to the reasoning mode in D_2 are as follows showing and examples are shown in Table 12:

1. **Latent Reasoning**
   When the model engages in latent reasoning, it tends to prioritize external information. Even when the model's memory is accurate, it can be significantly influenced by misleading context.
2. **Factual Shortcut**
   Even if the model has already undergone factual shortcuts, the base fact may still trigger another shortcut. The model exhibits reduced robustness when facing disruptive inputs.

For D_4, there could be many possible reasons, and it is also related to the model's generalization and generation capabilities.

H More Discussions about Our Experiments

H.1 Influence of Object Popularity on Ranking Experiments

In this section, we further discuss the potential impact of object popularity on the results of our ranking experiments. Object popularity refers to the frequency with which a particular object appears in large corpora or in diverse contexts, which could affect how the model ranks different objects when subjected to multi-hop reasoning tasks.

Our initial ranking experiment provided some insights, but we observed that object popularity could introduce biases that impact the model's ranking behavior. Specifically, objects with higher popularity may be ranked more highly, even when they do not semantically fit the context of a given question. This phenomenon could distort the model's decision-making process, particularly in multi-hop queries where the model must navigate through complex relations and intermediate entities.

We hypothesize that the model may rely on object popularity as a shortcut, especially in scenarios where the model is uncertain about the correct answer. The influence of object popularity may be especially pronounced when the model faces noisy or conflicting inputs, as it could prioritize familiar, frequently occurring objects over less common, but semantically relevant, entities.

To evaluate this influence in greater detail, we propose conducting additional ranking experiments, focusing on the following aspects:

- **Measuring Object Popularity:** To systematically assess the impact of object popularity, we will first measure it by analyzing large text corpora to obtain frequency statistics. Popularity will be assessed based on the frequency with which objects appear in various contexts, such as news articles, encyclopedias, and other publicly available sources.
- **Selecting Object Pairs with Similar Popularity:** To isolate the effect of popularity from other factors, we will select object pairs with similar popularity but differing semantic associations. For example, we might compare "Paris" (as a capital city) with "Einstein" (as a famous scientist). Both may have similar popularity, but their semantic contexts are distinct, making them suitable for testing the role of object popularity in ranking.

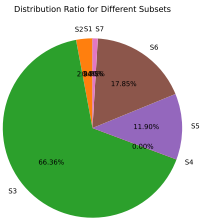

Figure 10: Proportion of the object in each subset.

- **Conducting Comparative Experiments:** We will replace objects in the ranking experiments and analyze whether the model's behavior is influenced by the popularity of the object. Specifically, we will examine if the POC/SOC (Popularity of Object/Subject of Change) effect persists when objects are swapped in different scenarios.

The results of these experiments will help us understand how object popularity interacts with the model's reasoning mechanisms, particularly in the context of multi-hop reasoning tasks. If object popularity proves to have a significant influence, we will explore strategies to mitigate its effects, ensuring more accurate and contextually relevant rankings in future models.

H.2 Object-related Subset Issue

In this section, we address an important aspect of our experimental design: the **Object-related Subset Issue**. Our investigation into multi-hop reasoning required us to define subsets based on the presence and relevance of objects, ensuring the analysis remained focused on the model's reasoning modes rather than unrelated variables. We divided the data into seven subsets to capture different interactions between the objects and bridge entities.

H.2.1 1. Proportion of the Subset Containing the Object

To study the object-related subset distribution, we created seven subsets, each containing different combinations of the object's involvement across various relational contexts. These subsets were designed to isolate the role of the object in multi-hop reasoning, focusing on how object relevance influences the overall model behavior. The proportion of the object in each subset was calculated based on the dataset, as shown in the Figure **??**: This distribution shows the varying degrees of object

influence across subsets, with the majority of the object data residing in subset $S3$, where objects play a critical role in the reasoning process.

H.2.2 2. Subset Choice: Why Not $Ab - As - Ao + Abos$?

In multi-hop reasoning, we specifically chose subsets that maintain the focus on **bridge-related attributes** rather than those that complicate the subject's enrichment process. Our key insight is that the enrichment of the subject position is crucial for reasoning, and we aim to exclude unnecessary interferences, such as unrelated objects, that do not contribute to this process.

Choosing the subsets that isolate the bridge-related attributes allows us to **control for variables** that could distract from our main objective, i.e., analyzing how the model encodes bridge information during reasoning. This approach ensures that we focus on the most relevant components in multi-hop tasks.

H.2.3 3. Adjustments and Testing

To ensure the object's effect on the reasoning process was appropriately considered, we conducted further testing by adjusting the weight of the object across subsets. This adjustment did not significantly alter the model's performance or reasoning behavior, confirming that the object's influence, when controlled, does not overshadow the bridge-related reasoning mechanisms.

Additionally, we conducted experiments using a **smoothing adjustment strategy** for the logits associated with the object, which helped to prevent excessive interference from the object's presence. Our testing results showed that the model's output was more influenced by the **attribute information** rather than by the object itself.

H.2.4 4. Conclusion

The object-related subset issue highlights the importance of careful dataset design when investigating multi-hop reasoning in large language models. By isolating the effects of objects and focusing on bridge-related reasoning, we ensure that the experiments accurately reflect the internal mechanisms at play. The use of subsets where the object plays a controlled role allows us to better understand how the model encodes and utilizes multi-hop information, paving the way for more robust future investigations into reasoning patterns and model interpretability.

I Extending ARR to Multi-hop Reasoning

The ARR metric is inherently model-agnostic and task-agnostic, relying solely on semantic correlations in hidden states. This makes it naturally extensible to more complex reasoning scenarios beyond two-hop tasks. Here, we demonstrate its application to three-hop reasoning.

I.1 ARR Application to Three-Hop Reasoning

We analyze a three-hop reasoning example to demonstrate ARR's extensibility:

Question: "What is the capital city of the country where the scientist who discovered radium was born?"

- First-hop: s = "scientist who discovered radium" $\rightarrow$ b_1 = "Marie Curie"
- Second-hop: b_1 = "Marie Curie" $\rightarrow$ b_2 = "Poland" (country of birth)
- Third-hop: b_2 = "Poland" $\rightarrow$ o = "Warsaw" (capital city)

Table 13 shows the structure of ARR computation across multiple reasoning hops:

Hop	Bridge (b)	Object (o)	ARR Computation
1-hop	Marie Curie	Poland	ARR(b_1, o_1)
2-hop	Poland	Warsaw	ARR(b_2, o_2)
Final	Warsaw	—	For final prediction

Table 13: Structure of ARR computation for multi-hop reasoning tasks.

Table 14 shows the ARR values measured across different layers:

Layer	First Hop			Second Hop		
	logit(b_1)	logit(o_1)	ARR_1	logit(b_2)	logit(o_2)	ARR_2
50	0.32	0.41	-0.09	0.25	0.20	+0.05
55	0.45	0.42	+0.03	0.33	0.29	+0.04
65	0.61	0.47	+0.14✓	0.54	0.36	+0.18✓

Table 14: ARR values where b_1=Marie Curie, o_1=Poland, b_2=Poland, and o_2=Warsaw.

I.2 Insights from Multi-hop ARR Analysis

Several important patterns emerge from our extended ARR analysis:

1. **Hierarchical Processing**: The transition from negative to positive ARR values demonstrates that the model gradually shifts from shortcut behavior to structured reasoning as information progresses through layers.

2. **Layer Specialization**: Even in large models, reasoning steps tend to concentrate in the mid-to-late layers (beyond layer 50 in our example), suggesting that multi-hop reasoning is not uniformly distributed but emerges prominently in later processing stages.

3. **Step-wise Verification**: ARR can effectively track each step of a multi-hop reasoning chain (ARR_1, ARR_2, etc.), allowing researchers to pinpoint where and how specific reasoning steps occur within the model.

I.3 Future Directions for ARR in Complex Reasoning Tasks

The ARR methodology can be naturally extended to analyze more complex reasoning patterns:

1. **N-hop Generalization**: ARR can be computed recursively for each hop in arbitrarily long reasoning chains, providing a consistent measurement framework across reasoning complexity levels.

2. **Reasoning Graph Analysis**: For tasks with branching reasoning paths, multiple ARR measurements can track parallel reasoning processes and identify which paths most influence the model's final decision.

3. **Cross-architectural Comparisons**: As a model-agnostic metric, ARR enables standardized comparison of reasoning mechanisms across different model architectures and scales, potentially revealing how architectural choices impact reasoning capabilities.

By applying ARR to more complex reasoning scenarios, we can develop a more comprehensive understanding of how transformer-based language models implement multi-step reasoning.

Prompt	Layer	Top k tokens
The country of citizenship of the spouse of Henry Clifford, 2nd Earl of **Cumberland** is	5	'ord', 'ords', 'ORD', 'lord', 'Nord', 'Gordon', 'Lord', 'Bruno', 'Jord', 'orney', 'Borg', **'Ford'**, 'orde', 'ardon', **'Leonard'**, 'ordon', 'Jordan', 'org', 'ardo', 'Lincoln', 'Cord', 'Meyer', 'order', 'laravel', 'orden', '', 'afford', 'Cleveland', '', '', 'Clark', 'orm', 'odor', 'Paul', 'dorf', 'wd', 'Oliver', 'üss', 'ald', 'eras', 'org', 'ford', 'Morris', 'Oriental', '', 'revision', 'örd', 'Carter', 'uv', 'med', 'roid', 'icy', 'longest', 'iegel', 'Vincent', 'cord', 'abil', 'bord', 'afka', 'olk', 'anda', 'thur', 'intendo', 'igneur'
	15	'enson', 'l**and**', 'Stanley', 'endorf', 'ington', 'Mountains', 'inton', 'cki', 'eston', 'indeed', 'emberg', 'department', 'industrial', 'ardin', 'yard', 'enty', 'ena', 'ley', 'öv', 'ember', '', 'andon', 'dimensional', '**England**', 'Mountain', 'ani', 'burgo', '', 'Pakistan', 'Thompson', 'eland', 'Holland', 'specification', 'founder', '', 'Williams', 'mouth', 'ṇ', 'ional', 'Prince', 'inten', 'ruck', 'numbers', 'heid', 'javascript', '', 'oux', 'entic', 'Kent', 'jal', 'highly', 'Sher', 'burg', 'Richmond', 'achi', 'snow', 'ęd', 'bland', 'Canada', 'fection', 'rias', 'anson', 'abb', 'yman', 'agar', 'burgh', '', 'Connecticut', 'ham', 'Robinson', 'stick', 'actor'
	25	**'land'**, 'enson', 'bury', 'ington', 'Stanley', 'gren', 'anton', 'eston', 'eland', 'endorf', 'mouth', 'chester', 'leton', 'emberg', 'Franklin', 'Maryland', **'England'**, 'dale', 'Duke', 'Department', 'composition', 'fly', '', 'borough', 'ardin', 'Edinburgh', 'igny', 'irmingham', 'Scotland', 'orton', 'burg', 'folk', '', 'inden', 'ivan', '', 'Russell', 'cki', 'hardt', 'York', 'hausen', 'beck', 'Lincoln', 'stone', 'enberg', '', 'lease', 'rei', 'anson', 'inton', 'County', 'inson', 'unt', 'ortheast', 'insen', 'ansen', 'olk', 'Braun', 'mole', 'District', 'heim', '', 'ols', '', 'ström', 'bol', 'lyn', 'Leopold', 'ford', '', 'founder', 'aki', 'ama', 'onian', 'personally'

Table 9: Top-k tokens extracted from different layers for the first subject.

Prompt	Layer	Top k tokens
The capital of the country with **the Pyramids of Giza** is	5	'iza', **'Egypt'**, 'izar', 'iz', 'gypt', 'izia', 'airo', 'Peru', 'itza', 'isa', 'tomb', 'za', 'Elis', 'ixa', 'isi', 'icia', 'aza', 'ifa', 'ja', 'Gia', 'Jerusalem', '', 'Lis', 'Gaz', 'tick', 'Iz', 'aris', 'izo', 'iso', 'IZ', 'zeta', 'hausen', 'ya', 'amaz', 'nitz', 'itzer', 'cca', 'inta', 'hoff', 'cian', 'zyk', 'ixon', 'biz', '', 'izz', 'Jung', 'ocia', 'zza', 'pc', 'ira', 'Roma', '', '', 'izations', 'ancient', 'isie', 'arte', '', 'observ', 'baz', 'aka', 'ization', 'izza', 'Paris', 'existed', 'yard', 'ka', '', 'ba', 'zo', 'ysz', 'osi', '', 'Krak', 'lava', 'zien', 'jar', 'aga', 'zat', 'endl'
	15	**'Egypt'**, 'iza', 'gia', 'gypt', 'ica', 'Peru', **'airo'**, 'izia', 'isie', 'Janeiro', 'isi', 'auff', 'igo', '', 'Grey', 'bia', 'Jordan', '', 'Houston', 'Miami', 'uez', '', 'fica', 'unda', 'phrase', 'icia', 'uga', 'ptic', 'onian', 'Africa', 'Brazil', 'hoz', 'inta', 'ids', 'Alexand', 'eda', 'izza', 'itzer', 'Singapore', 'anska', '', 'projects', 'bi', 'raz', 'gender', 'era', 'yrus', '', 'fico', 'gray', 'stones', 'ña', 'indi', 'python', 'Kent', 'effic', '', 'shoulder', 'antine'
	25	'icians', 'documents', '</s>', 'imin', '', '', '', '', 'vin', 'onian', 'ification', 'Gib', 'istan', 'Gil', 'Metropolitan', '', 'mouth', 'ford', 'inda', '', 'allow', 'agon', 'scribe', 'ou', 'isher', 'ey', 'ital', 'inclus', 'allen', 'Hamilton', 'Valley', 'Wayne', 'iza', 'ilities', 'validate', 'ingham', '', 'Gia', 'assigning', 'andra', 'Lewis', 'aris', 'Jordan', 'Roberts', 'burg', '', 'Ryan', 'WD', 'Gray', 'distance', 'ette', 'ila', 'simultaneously', 'gress', 'hire', 'oz', 'anti', 'angel', 'stone', 'ara', 'connections', '', 'generalized', 'illa', 'Foundation', 'Gran', 'iche', 'Program', 'icular', 'achi', 'general', 'regression', 'stein', 'ayer', '', 'ptic'

Table 10: Top-k tokens extracted from different layers for the second subject.

Scenario	Base Fact (c-)	Base Question (D+->D-)
ARR > 0 (latent reasoning)	Callum McManaman was born in France	In which borough was Callum McManaman born?
ARR < 0 (factual shortcut)	Einstein's spouse is Elizabeth.	Where was Einstein's spouse born?

Table 12: Examples in D2 (to be expanded in Appendix).

Figure 11: 7B Window 1

Figure 12: 13B Window 1

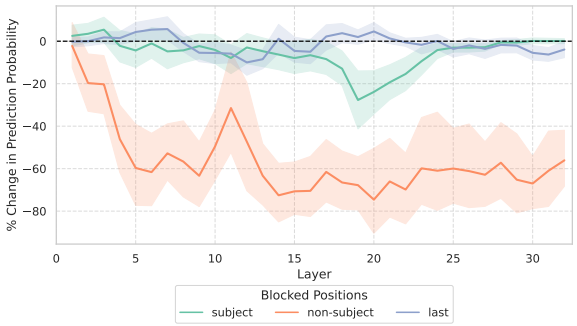

Figure 13: 7B Window 5

Figure 14: 13B Window 5

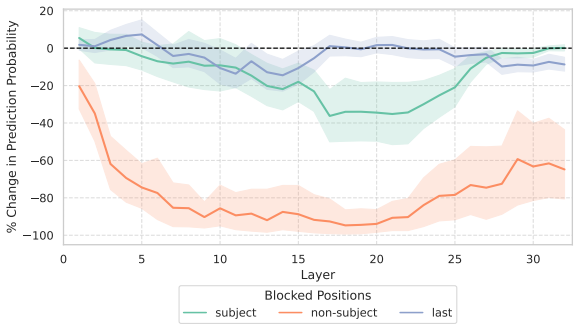

Figure 15: 7B Window 10

Figure 16: 13B Window 10

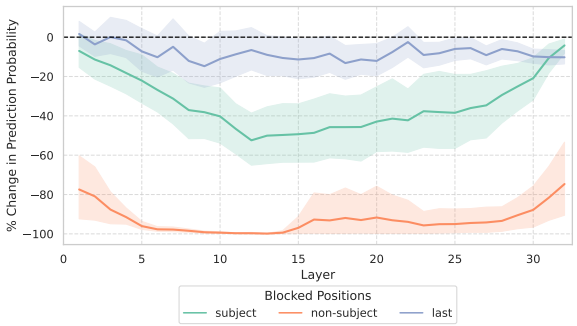

Figure 17: 7B Window 20

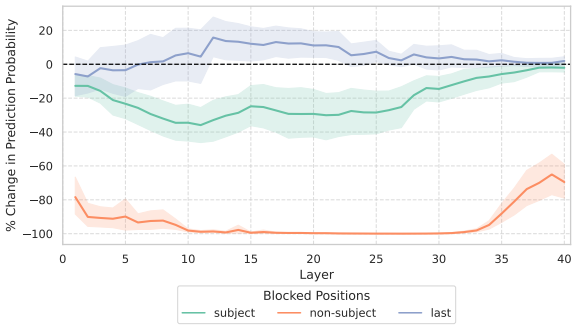

Figure 18: 13B Window 20

Figure 19: Combined results for different windows (7B and 13B).