Pierce the Mists, Greet the Sky: Decipher Knowledge Overshadowing via Knowledge Circuit Analysis

Anonymous ACL submission

Abstract

Large Language Models (LLMs), despite their remarkable capabilities, are hampered by hallucinations. A particularly challenging variant, **knowledge overshadowing**, occurs when one piece of activated knowledge inadvertently masks another relevant piece, leading to erroneous outputs even with high-quality training data. Current understanding of overshadowing is *largely confined to inference-time observations, lacking deep insights into its origins and internal mechanisms during model training*. Therefore, we introduce **PHANTOMCIRCUIT, a novel framework designed to comprehensively analyze and detect knowledge overshadowing**. By innovatively employing knowledge circuit analysis, PHANTOMCIRCUIT dissects the internal workings of attention heads, tracing how competing knowledge pathways contribute to the overshadowing phenomenon and its evolution throughout the training process. Extensive experiments demonstrate PHANTOMCIRCUIT 's effectiveness in identifying such instances, offering novel insights into this elusive hallucination and providing the research community with a new methodological lens for its potential mitigation.

1 Introduction

Large Language Models (LLMs) have witnessed explosive growth in recent years, demonstrating remarkable capabilities across a multitude of domains, including natural language understanding, generation, reasoning, and even cross-modal tasks (Chang et al., 2024; Zhao et al., 2024; Yan et al., 2025a, 2024a; Xun et al., 2025). Their proficiency has catalyzed transformative advancements in various applications. However, a persistent challenge that tempers their widespread adoption and reliability is the phenomenon of hallucination. Broadly, hallucinations refer to instances where models generate content that is factually incorrect, nonsensical, or unfaithful to the provided source context, despite

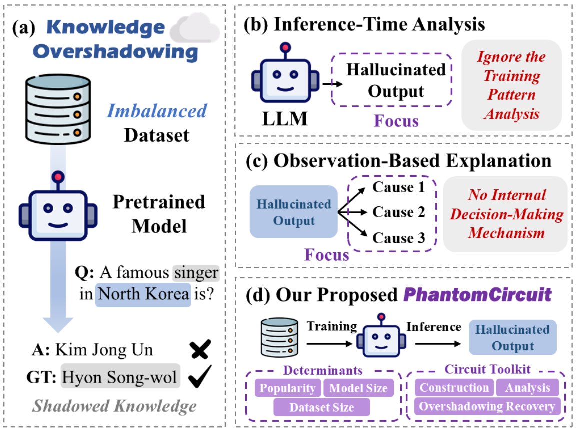


Figure 1: Illustrative comparison of previous research with inference-time analysis (b) and observation-based explanation (c) vs our proposed PHANTOMCIRCUIT (d) on knowledge overshadowing (a).

appearing coherent and fluent (Rawte et al., 2023; Chakraborty et al., 2025; Wang et al., 2025).

While substantial research has been dedicated to understanding the causes and detection of general hallucinations, a specific variant known as "**knowledge overshadowing**" warrants deeper investigation (Zhang et al., 2024b, 2025a). This phenomenon is particularly perplexing because it can manifest even when models are trained on high-quality, meticulously curated pre-training corpora. Current understanding, primarily derived from inference-time observations, characterizes overshadowing as a scenario where, for a given query, one piece of activated knowledge inadvertently "overshadows" another relevant knowledge. This interference ultimately biases the model's reasoning process, leading to a hallucinatory output, as illustrated in Figure 1 (a).

Nevertheless, existing explorations into knowledge overshadowing suffer from notable limitations. ❶ They *predominantly focus on inference-time analysis*, as shown in Figure 1 (b). While valuable for identifying the occurrence of overshadowing, such observations offer a surface-level understanding and often fall short of elucidating how these detrimental patterns are learned during the training phase. ❷ The explanations for overshadowing are often *speculatively inferred from these observational outcomes* rather than being rigorously investigated through dedicated interpretability tools that can probe the model's internal decision-making mechanisms, as shown in Figure 1 (c). Consequently, a more comprehensive analytical framework is imperative to dissect this phenomenon from its origins to its manifestation.

To bridge this gap, we introduce **PhantomCircuit, a novel framework designed to comprehensively analyze and detect the knowledge overshadowing phenomenon**. Specifically, PhantomCircuit facilitates an in-depth examination of the evolution of overshadowing hallucinations throughout the training process, correlating their emergence and prevalence with core factors such as knowledge popularity, model size, and dataset size. Then, by leveraging knowledge circuit analysis as a key interpretability technique, we aim to trace the flow of information and the formation of knowledge representations within attention heads, thereby uncovering the internal mechanisms giving rise to overshadowing. Furthermore, we propose to optimize the number of edges within these circuits, thus alleviating the knowledge overshadowing. As illustrated in Figure 1 (d), our overall work aims to provide a clearer, mechanistic understanding and potential strategies for this elusive type of hallucination.

Our contributions can be summarized as follows:

- We introduce PhantomCircuit, the first comprehensive framework designed to systematically analyze and detect knowledge overshadowing, delving into its mechanistic nature and evolution throughout model training.

- We pioneer the use of knowledge circuit analysis to dissect the internal workings of attention heads, specifically elucidating how competing knowledge pathways contribute to the overshadowing phenomenon.

- We conduct extensive experiments to demonstrate PhantomCircuit 's efficacy in detecting knowledge overshadowing, offering novel insights and a new methodological lens for the research community.

2 Related Work

2.1 Hallucination Detection

Factuality hallucination detection, which aims to evaluate whether the output of LLMs aligns with real-world facts, typically involves either external fact-checking or internal uncertainty analysis (Huang et al., 2025; Dang et al., 2024; Zheng et al., 2024; Zou et al., 2024; Zhou et al., 2024; Zhu et al., 2024). For instance, FACTSCORE (Min et al., 2023) decomposes a generation into atomic facts and calculates the proportion that are supported by reliable knowledge sources. FACTOOL (Chern et al., 2023), on the other hand, integrates multiple tools such as Google Search and Google Scholar to gather external evidence and assess the factuality of generated content. In contrast, methods like Chain-of-Verification (Dhuliawala et al., 2023), probability-based assessments (Kadavath et al., 2022; Zhang et al., 2024a), and uncertainty estimation approaches (Varshney et al., 2023; Yao et al., 2023; Luo et al., 2023) rely on LLMs' internal parametric knowledge or uncertainty signals to predict potential hallucinations. Among these efforts, knowledge overshadowing (Zhang et al., 2025a) offers a novel perspective by modeling hallucination behavior from the perspective of knowledge representation, providing an efficient strategy for proactive prevention.

2.2 Knowledge Circuit Analysis

In the context of mechanistic interpretability (Rai et al., 2024; Huo et al., 2024; Huang et al., 2024a), computations in Transformer-based language models are viewed as a connected directed acyclic graph encompassing components such as MLPs and attention layers (Syed et al., 2023; Conmy et al., 2023; Huang et al., 2024b). A **circuit** refers to a sparse computational subgraph that significantly influences the model's behavior on a specific task (Olah et al., 2020; Elhage et al., 2021; Wang et al., 2022). Building on this, Yao et al. (2024) introduce the concept of **knowledge circuits**, hypothesizing that cooperation among components reveals implicit knowledge in LLMs. Further, Ou et al. (2025) explore how such circuits evolve during continual pre-training, providing insights into knowledge acquisition. To enable effective knowledge editing, CaKE (Yao et al., 2024) proposes a Circuit-aware Knowledge Editing method that guides models to activate modified knowledge and form new reasoning pathways. In this paper, we analyze the phenomenon of knowledge overshadowing through the lens of knowledge circuits, contributing new perspectives to LLM hallucination detection.

See more related work in Appendix A.1.

3 Methodology

This section first defines the knowledge overshadowing phenomenon and its quantitative evaluation. Subsequently, we detail the PHANTOMCIRCUIT framework, which encompasses methods for analyzing its training dynamics and for constructing and analyzing knowledge circuits to understand its internal mechanisms[1].

3.1 Knowledge Overshadowing

Knowledge overshadowing refers to a specific type of hallucination where *less prevalent, subordinate knowledge* is suppressed by *high-frequency, dominant knowledge* when both are associated with *common background knowledge* (Zhang et al., 2025a).

Let X_{dom} denote dominant knowledge entities and X_{sub} denote subordinate knowledge entities, both potentially co-occurring with background knowledge X_{bg}. The core idea is that a strong learned pattern $X_{dom} \leftrightarrow X_{bg}$ can "overgeneralize" where the model primarily associates X_{dom} with X_{bg}.

Consequently, when the model encounters X_{sub} with X_{bg}, denoted as $X_{sub} \leftrightarrow X_{bg}$), it may erroneously favor outputs related to X_{dom} due to the stronger $X_{dom} \leftrightarrow X_{bg}$ pattern.

3.1.1 Knowledge Overshadowing Occurrence

When the input prompt is a knowledge pair composed of background knowledge and dominant knowledge, denoted as $P_{dom} = (X_{bg}, X_{dom})$, and the model correctly generates the answer corresponding to the dominant knowledge, denoted as Y_{dom}, we consider this outcome, represented by the pair (P_{dom}, Y_{dom}), as a successful recall of dominant knowledge.

When the input prompt is P_{sub}, but the model wrongly generates Y_{dom}, knowledge overshadowing occurs for this query-response instance, resulting in the (P_{sub}, Y_{dom}).

3.1.2 Quantitative Indicators

Let N_{dom} and N_{sub} be the number of instances of P_{dom} and P_{sub} in the training dataset, respectively. The dataset Z comprises a subset Z_{dom} containing N_{dom} instances of P_{dom}, and a subset Z_{sub} containing N_{sub} instances of P_{sub}.

During autoregressive generation tasks performed by the model, let M_{sub} be the number of times when overshadowing instances (P_{sub}, Y_{dom}) occur , and M_{dom} be the number of times (P_{dom}, Y_{dom}) occurs. Then, we can define the absolute extent of the knowledge overshadowing effect, the Absolute Overshadowing rate ($\mathcal{AO}$) and calculate it using M_{sub} and N_{sub},

$$\mathcal{AO} = p(Y_{dom}|P_{sub}) = \frac{M_{sub}}{N_{sub}}. \tag{1}$$

To account for the model's inherent performance and potential noise affecting the overshadowing rate, we also introduce R_{dom} for dominant knowledge inputs:

$$R_{dom} = p(Y_{dom}|P_{dom}) = \frac{M_{dom}}{N_{dom}}, \tag{2}$$

which represents the recall rate for P_{dom} query-response instances.

The Relative Overshadowing rate ($\mathcal{RO}$) is then defined as:

$$\mathcal{RO} = \frac{\mathcal{AO}}{R_{dom}} = \frac{p(Y_{dom}|P_{sub})}{p(Y_{dom}|P_{dom})}. \tag{3}$$

3.1.3 Overshadowing Influence Factors

Knowledge Popularity (P) is the fundamental cause of the knowledge overshadowing phenomenon and serves as its primary influencing factor. P is defined as the ratio of the capacities of Z_{dom} to Z_{sub}, thus $P = N_{dom}/N_{sub}$.

Model Size (M), referring to the number of model parameters, also impacts knowledge overshadowing. A larger M generally implies stronger generalization capabilities, causing the model to rapidly generalize the $X_{dom} \leftrightarrow X_{bg}$ to instances involving $X_{sub} \leftrightarrow X_{bg}$ and exacerbating overshadowing.

In addition to the factors mentioned in (Zhang et al., 2025a, 2024b), aiming to analyze the dynamic evolution of knowledge overshadowing during the training process, we extend our consideration to total number of tokens, the **Dataset Size (D)** in the training set. The **average loss proportion of subordinate knowledge $\mathcal{LP}$ within an epoch** is defined as $\mathcal{LP} = loss(P_{sub})/total\ loss$, which also relates to the overshadowing.

[1] Our code will be available upon acceptance.

3.2 Analysis Framework PHANTOMCIRCUIT

Our proposed knowledge circuit-based overshadowing analysis framework involves the overshadowing dynamics analysis during training and circuit-based internal mechanism analysis.

3.2.1 Overshadowing Dynamic Analysis

Our framework provides a novel dynamic analysis of knowledge overshadowing during model training. By manipulating P, M, and D, we monitor $\mathcal{RO}$ across epochs. We focus on identifying the onset, duration, and recovery stages of overshadowing to understand their modulation by P, M, and D. Recognizing P as a key factor, we also explore how $\mathcal{LP}$ co-evolves with $\mathcal{RO}$ under these variables, aiming to uncover the role of $\mathcal{LP}$ in explaining overshadowing dynamics.

3.2.2 Knowledge Circuit Construction

We construct the knowledge circuit, a sparse computational subgraph $C \subseteq G$, where $G = (V, E)$ is the directed acyclic graph representation of LLMs, V encompasses input embeddings, attention heads, MLP layers, and output logits, E represents the information flow between these components. Our goal is to identify a subgraph C that is critical for recognizing the key component of given input prompt, particularly in knowledge overshadowing, is $\{X_{dom}, X_{sub}\}$, the difference between P_{dom} and P_{sub}. The adapted construction methods is similar to edge attribution patching (Conmy et al., 2023), which involves:

Paired inputs. For a given X_{bg}, we create a pair of input prompts, the clean input P_{sub} corresponding to our expected output Y_{sub} in the overshadowing case, and the corrupt input P_{dom} serving for the contrasting mentioned below.

Activation difference calculation. After running this pair of inputs through the model, we calculate the difference in activation values $\Delta A(v_p), \Delta A(v_c)$ for the parent node v_p and child node v_c of each edge under these distinct inputs, $\Delta A(v_p) = A_{clean}(v_p) - A_{corrupt}(v_p)$, where $A_{clean}(v_p)$ and $A_{corrupt}(v_p)$ are the activation of parent node when P_{sub} as clean input and P_{dom} as corrupt input.

Edge score calculation. The importance $S(e)$ of an edge $e = (v_p, v_c)$ is scored based on how patching v_p's activation (using $\Delta A(v_p)$) influences a metric $\mathcal{M}$, which assesses the model's ability to correctly output Y_{sub} rather than Y_{dom} for P_{sub} inputs. Using the Integrated Gradients, $S(e)$ is approximated as:

$$S(e) \approx Exp\left[\Delta A(v_p) \cdot \frac{\partial \mathcal{M}(Y_{sub}|P_{sub})}{\partial A(v_p)}\right]. \quad (4)$$

Circuit construction. Edges with scores $|S(e)|$ below a threshold τ are pruned. The remaining subgraph forms the knowledge circuit C. See more details about construction in Appendix B.1

3.2.3 Circuit-based Analysis

PHANTOMCIRCUIT mainly focuses on the attention heads in C and follows these steps:

Node Attention Analysis. We identify high attention heads within C by examining their attention scores and patterns, specifically their focus on $\{X_{dom}, X_{sub}\}$.

Circuit Structure Analysis. We then trace the information flow of these high attention heads by identifying their parent and child nodes to understand their structural role. Nodes consistently retained in circuits built with different thresholds τ are also analyzed as key components.

Layer-wise Logit Evolution. Finally, using logit lens (nostalgebraist, 2020), we inspect the evolving output logits at layers associated with key nodes. This validates if their captured information contributes to the model's prediction as expected.

3.2.4 Circuit-based Overshadowing Recovery

Inspired by the goal of knowledge circuits to maximize sensitivity to the distinguishing features between X_{dom} and X_{sub}, we propose a circuit-based method to alleviate overshadowing. This involves optimizing the circuit structure by tuning the edge pruning threshold τ to obtain an optimal circuit, C_{opt}. The optimization is formulated as:

$$\tau_{opt} = \arg\max_{\tau} \mathcal{M}(C_{opt}(\tau), P_{sub}, Y_{sub}), \quad (5)$$

where $\mathcal{M}$ measures the circuit's ability to distinguish $\{X_{dom}, X_{sub}\}$, This results in a $C_{opt}(\tau_{opt})$ which is expected to mitigate or eliminate the overshadowing effect for specific input prompts.

To automate the overshadowing recovery process and extend its applicability, we simplify the relative pointwise mutual information (R-PMI) method from (Zhang et al., 2025a, 2024b) to identify X_{sub} within P_{sub}. First, we obtain the top-k next-token candidates, $V_{top}(P_{sub})$, by feeding P_{sub} to the model. Then, we iteratively generate contrastive prompts P'_{sub} by masking (in our implementation, by deleting) each token X'_{sub} from P_{sub}. For each P'_{sub}, we acquire its top-k candidates $V_{top}(P'_{sub})$. The R-PMI for each token y_i in the intersection $V_{top}(P_{sub}) \cap V_{top}(P'_{sub})$ is calculated as:

$$R\text{-}PMI_i = \log \frac{p(y_i|P_{sub})}{p(y_i|P'_{sub})}. \quad (6)$$

Then, we sum the negative R-PMI values to get $S_{R\text{-}PMI} = \sum \min(R\text{-}PMI_i, 0)$. The X'_{sub} yielding the minimum $S_{R\text{-}PMI}$ is identified as X_{sub}.

Furthermore, the Y_{sub} is determined as the token y_i from $V_{top}(P_{sub})$ whose average rank improves most significantly when non-subordinate components X'_{sub}, often the X_{bg}, is masked. Y_{dom} is identified as y_i that has the highest average rank across all P'_{sub}. For circuit construction, X_{dom} within P_{dom} is replaced by a generic placeholder like "something", as shown in Figure 5. Combining this streamlined approach enables broader application of our knowledge circuit-based overshadowing recovery. See more details in Appendix B.2

4 Experiment

4.1 Experiment Setup

4.1.1 Dataset

Synthetic dataset. To investigate the dynamic characteristics and influencing factors of the $\mathcal{RO}$ during training under controlled conditions,and to minimize the complexities and semantic relationships inherent in natural language, we construct a synthetic dataset and train models from scratch.

Follow (Zhang et al., 2025a), we first fix the length of X_{bg} at 4 tokens and the lengths of $X_{dom}, X_{sub}, Y_{dom}, Y_{sub}$ at 1. For the dataset size (D), we experiment with 0.26 (D = 0.26M), 2.6 (D = 2.6M) and 26 million tokens (D = 26M). For popularity (P), we set P values as 5, 25, and 100.

Then, for a specific combination of P and D, there are several distinct groups to achieve the target D. Each group comprises P distinct X_{dom} and a single Y_{dom} for dominant knowledge, one X_{sub} and Y_{sub} for subordinate knowledge. The X_{bg} is shared. Thus, there are P unique $\{P_{dom}, Y_{dom}\}$ and one $\{P_{sub}, Y_{sub}\}$ in a group. All the tokens are randomly sampled. See more details of our dataset in Appendix C.

Finetuning Dataset. We construct a finetuning dataset to evaluate circuit-base overshadowing recovery method. Utilizing virtual knowledge, we preserve the natural language semantics while avoiding prior knowledge editing that could interfere with the natural occurrence of overshadowing. For this dataset, we set P = 5 and D = 1M. The Appendix D shows the dynamics analysis based on finetuning dataset.

4.1.2 Models Evaluated

We employ models from the Pythia suite (Biderman et al., 2023), specifically: Pythia-70M, Pythia-410M, Pythia-1.4B, and Pythia-2.8B, corresponding to model sizes (M) of M = 70M, 410M, 1.4B and 2.8B, respectively. Tokens randomly sampled to build synthetic dataset is from Pythia tokenizer.

4.1.3 Training

A uniform learning rate of 10^{-5} and batch size of 16 are used for both dataset. Training is conducted on NVIDIA A800 GPUs.

4.1.4 Evaluation

To measure $\mathcal{RO}$, we randomly sample 500 P_{dom} and 500 P_{sub} prompts for evaluation after each training epoch. The $\mathcal{LP}$ is recorded within each epoch. The results are shown in Figures 2 and 3.

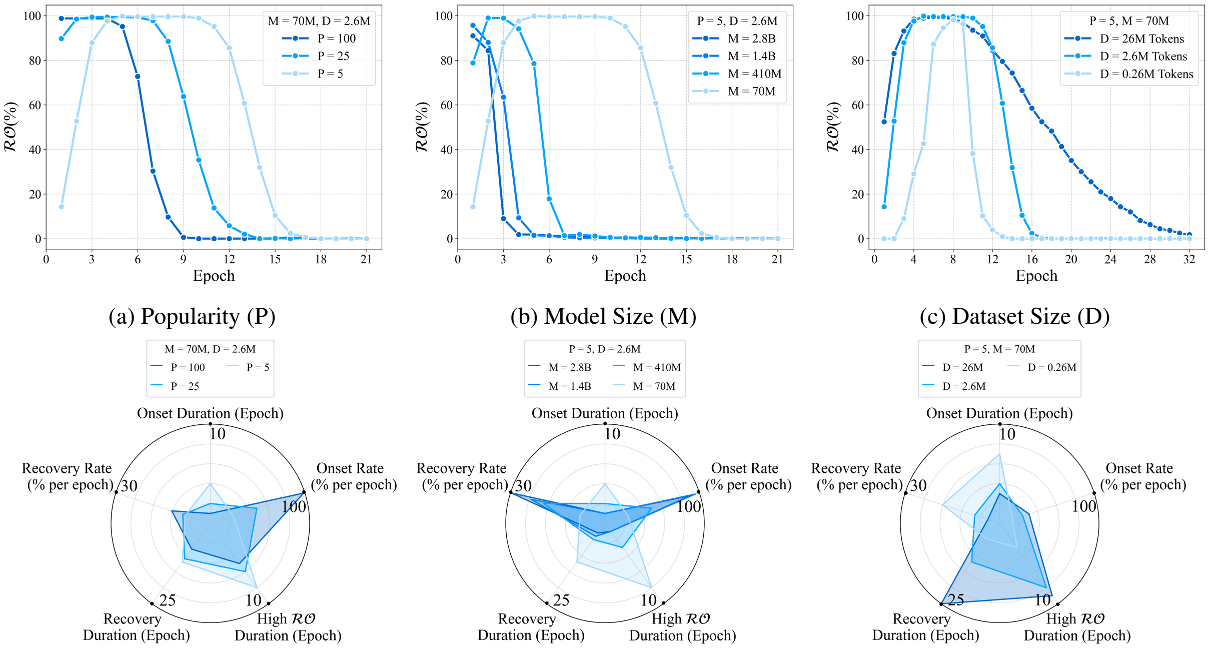
4.2 Main Result

Based on the experiments described above, using the circuit to analyze and optimize overshadowing, our investigation yields the following findings.

A higher value of P and M can lead to an earlier onset, shorter duration, and quicker recovery of the knowledge overshadowing. Distinctly, a larger D contributes to the earlier onset but also a slower recovery from overshadowing.

As shown in Figures 2a and 2d (M=70M, D=2.6M), increasing P (from 5 to 100) significantly shortens or even eliminates the onset phase. This is attributed to more prominent dominant patterns $X_{dom} \leftrightarrow X_{bg}$ being learned and generalized rapidly, even within the first epoch. The duration phase also decreases with higher P, because a larger P and a fixed D implies fewer groups of knowledge pairs, leading to less diversity of P_{sub}, allowing the model to learn all overshadowed P_{sub} and recover from overshadowing more quickly.

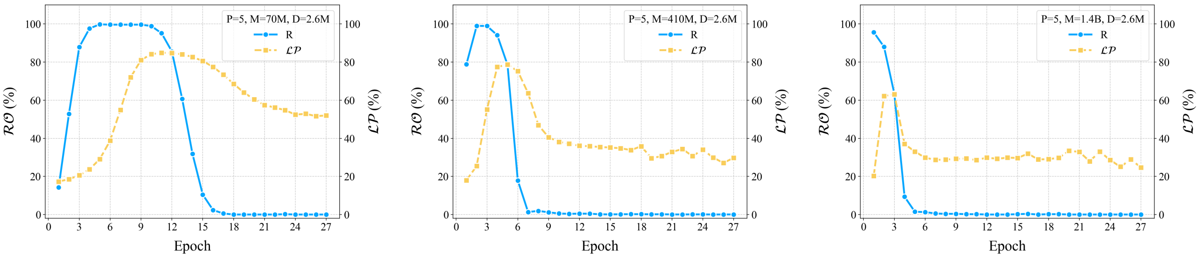
Figures 2b and 2e (P=5, D=2.6M) demonstrate that larger models (M) exhibit shorter or absent onset phases, indicating an earlier occurrence of knowledge overshadowing. This is due to the stronger generalization capabilities of larger models, leading them to quickly learn and overgeneralize dominant patterns. However, larger models also show a shortened duration phase and a rapid decline in $\mathcal{RO}$ during recovery, suggesting enhanced capacity to differentiate $\{X_{dom}, X_{sub}\}$, thus recovering faster despite earlier overshadowing.



(a) Popularity (P) (b) Model Size (M) (c) Dataset Size (D)

(d) Popularity (P) (e) Model Size (M) (f) Dataset Size (D)

Figure 2: (a) ∼ (c) show the dynamic variation of $\mathcal{RO}$ relating to P, M and D in model training phase. Higher Knowledge Popularity (P) and Model Size (M) tend to result in an earlier onset, shorter duration, and quicker recovery from knowledge overshadowing. In contrast, a larger Dataset Size (D) also leads to an earlier onset but is associated with a slower recovery phase. (d) ∼ (f) show the duration of onset stage, high $\mathcal{RO}$ (> 90%) stage as well as recovery stage, and $\mathcal{RO}$'s rate of change during the onset and recovery stages.



(a) M = 70M (b) M = 410M (c) M = 1.4B

Figure 3: The co-evolution between $\mathcal{RO}$ & $\mathcal{LP}$ in different M. Early high overall loss and low $\mathcal{LP}$ leads to intensive optimization of P_{dom} and $\mathcal{RO}$ rises up to near 100%. As training progresses, subordinate knowledge loss proportion ($\mathcal{LP}$) rises, shifting optimization focus to P_{sub} errors, initiating $\mathcal{RO}$'s recovery phase, validated across models.

In Figures 2c and 2f (P=5, M=70M) larger D leads to an earlier onset of overshadowing, with $\mathcal{RO}$ peaking sooner. This is because more data per epoch provides more iterations and exposure to the dominant pattern, accelerating its generalization. Conversely, the recovery phase is prolonged with larger datasets. The increased diversity of P_{sub} in larger datasets requires more epochs for the model to learn all instances and recover.

Notably, across various parameter combinations, $\mathcal{RO}$ often approaches 100% in the early training stages and finally recovers to 0%. We hypothesize this phenomenon stems from initial high-loss state, where optimization efforts disproportionately focus on reducing the larger loss contribution from P_{dom}. Our subsequent observations regarding $\mathcal{LP}$ corroborate this.

The dynamic nature of knowledge overshadowing arises from the co-evolution relationship between the $\mathcal{LP}$ and $\mathcal{RO}$. As depicted in Figure 3, in the early training stages, when the overall loss is high and the contribution from P_{sub} is small, the optimization process tends to concentrate on the P_{dom} and $\mathcal{RO}$ rapidly approaches 100%. However, as training progresses and $\mathcal{LP}$ begins to rise, eventually nearing its peak, P_{sub} takes a substantial portion of the remaining loss. At this juncture, the model's optimization efforts shift to focus on these errors from P_{sub}. This shift initiates the recovery phase, leading to a decline in $\mathcal{RO}$. Therefore, the insufficient optimization results in the overshadowing, which is consistent across varying M.

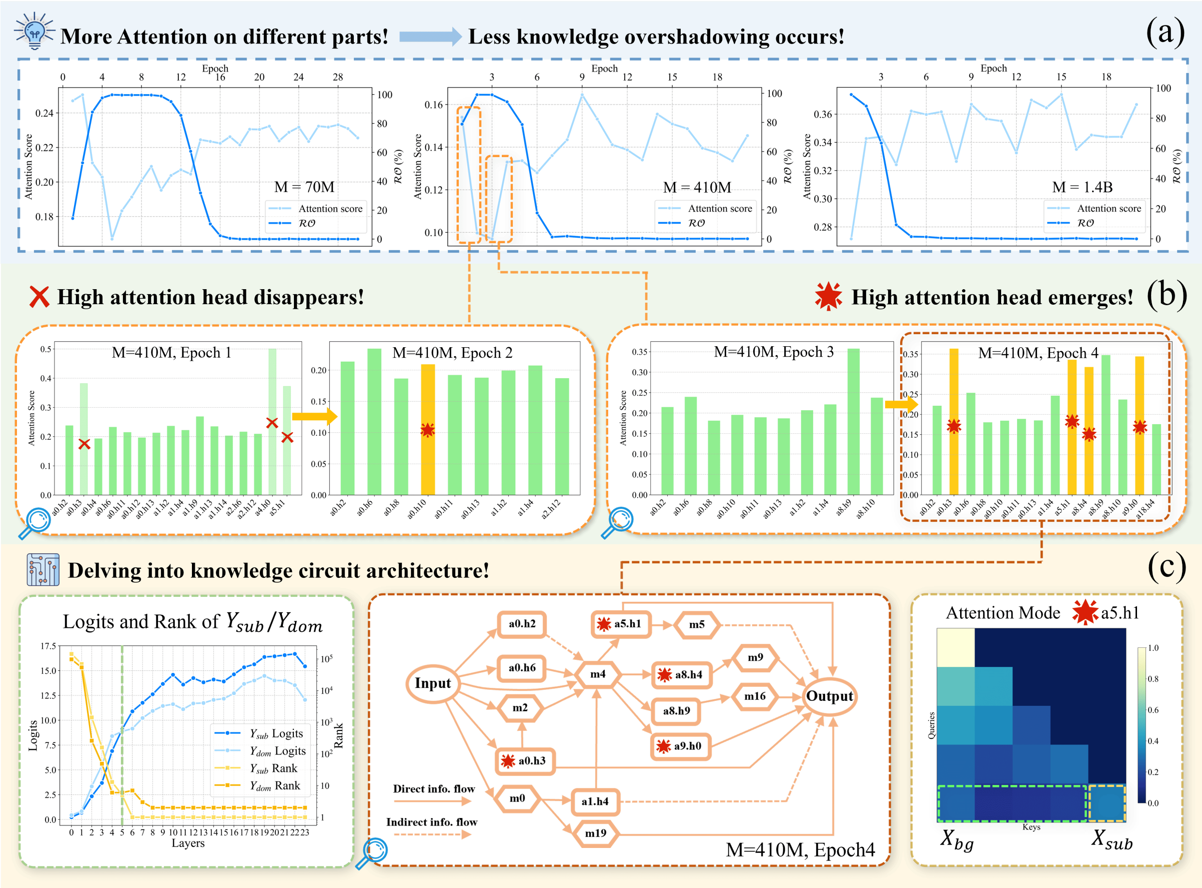


Figure 4: The circuit-based analysis of proposed PHANTOMCIRCUIT. (a) shows the average attention scores allocated to $\{X_{dom}, X_{sub}\}$ across epoch. (b) focuses on the onset and recovery phases and shows that the appear and disappear of high attention heads (attention score greater than 0.2) attribute to the reduce and raise of $\mathcal{RO}$. (C) shows the logits and ranks of Y_{sub}, Y_{dom} across layers. The main structure of circuit (with 400 edges totally) and attention mode show that high attention head a5.h1 and MLP layer m5 contribute to the juncture of Y_{sub} 's rank. Solid lines denote direct information flow, while dashed lines indicate indirect flow in circuit structure map.

The occurrence of overshadowing in pre-trained LLMs can be understood with dynamics analysis results above. Initially, large M and D promote the rapid generalization of $X_{dom} \leftrightarrow X_{bg}$ and overshadowing onset. The subsequent inadequate optimization of the P_{sub} is exacerbated by the sheer scale and diversity of training data, which brings prolonged overshadowing effect and exceeded sharp recovery effect from large M (e.g., trillion tokens for training LlaMa-2-7B).

The knowledge circuit's attentional allocation to differences between dominant and overshadowed knowledge inputs dictates the extent of knowledge overshadowing. Figure 4 (a) shows the variation of circuit's average attention scores on $\{X_{dom}, X_{sub}\}$ throughout the training phase. The higher average attention alleviates overshadowing, while lower attention exacerbates it.

Figure 4 (b) shows the circuit dynamics across the onset and recovery phases of knowledge overshadowing. These bar plots show the attention scores of individual attention heads in a specific epoch, indicating that when the $\mathcal{RO}$ declines, some attention heads, defined as high attention heads, exhibiting high focus on $\{X_{dom}, X_{sub}\}$ emerge within the circuit. The threshold for high attention is set at 0.2 for the length of X_{sub} is one fifth of the length of input prompt in synthetic dataset. Conversely, when knowledge overshadowing intensifies, a subset of these critical attention heads tends to disappear from the circuit.

Furthermore, by focusing on the circuit's internal mechanisms and structure within a specific epoch, as shown in Figure 4 (c), we leverage the circuit-based analysis of our proposed PHANTOMCIRCUIT. First, according to the logits and ranks of Y_{dom} and Y_{sub} across layers, the 5th layer is a juncture where the rank of Y_{sub} exceeds Y_{dom}. Subsequently, by focusing on the circuit structure, we identify the internal mechanisms driving this juncture. A high attention head, a5.h1 (layer 5), crucially channels information to the subsequent

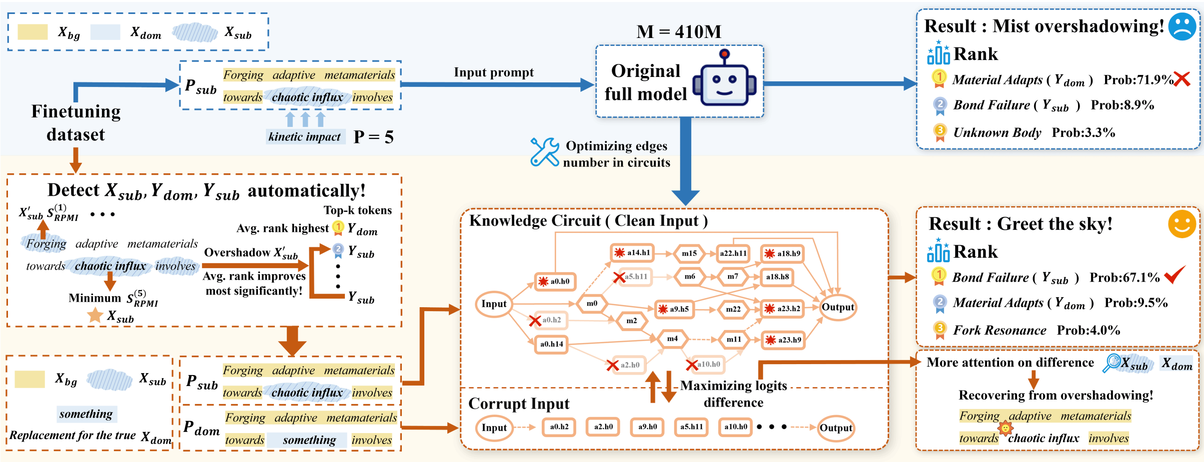


Figure 5: Overshadowing recovery via optimized circuit. In the finetuned model, $X_{dom} \leftrightarrow X_{bg}$ causes the original full model to incorrectly predict Y_{dom}. First, we detect the $X_{sub}, Y_{sub}, Y_{dom}$ automatically by calculating the minimum $S_{R\text{-}PMI}$. Then, optimizing the knowledge circuit by pruning edges to keep only key nodes enhances the attention on $\{X_{dom}, X_{sub}\}$, enabling the recovery from overshadowing and facilitating correct Y_{sub} prediction.

MLP layer, m5. The a5.h1 also appears to inherit its high attention state from earlier layers, mediated by the layer 4 MLP (m4), leading it to not only continue elevating the logits for Y_{sub} but also to attenuate the previously rapid growth of Y_{dom}'s logits, thereby facilitating the observed rank reversal. The attention map on the right confirms its high focus on $\{X_{dom}, X_{sub}\}$.

Knowledge circuit-guided optimization represents a promising strategy to mitigate the knowledge overshadowing effect. Figure 5 illustrates our findings. We first finetune the model on finetuning dataset. During the recovery phase, we randomly choose some P_{sub}. Feeding P_{sub} to the original full model (M = 410M), the prediction is still Y_{dom} instead of expected Y_{sub}, and shows a significant gap in probability as well.

We then detect the Y_{sub}, Y_{dom} and X_{sub} for chosen P_{sub} automatically, replace the X_{sub} of placeholder token "something". With these components, we obtain the P_{dom} as corrupt input and P_{sub} as clean input to construct a knowledge circuit. A golden section search algorithm is employed to determine the optimal number of edges for building C_{opt}. The optimized circuit structure map shows that some attention heads are pruned, which are often low attention heads, or exhibit no significant attentional pattern towards the core task-relevant information. Retained high attention heads are key to differentiating P_{dom} from P_{sub} which give significant attention to $\{X_{dom}, X_{sub}\}$. Some low attention heads also remain, implying that even in the circuit, processing background knowledge X_{bg} and linking it to the distinctive elements X_{sub} is crucial for correct inference. The performance of the optimized circuit C_{opt} is then evaluated by feeding it the clean input P_{sub}, while P_{dom} serves as the baseline for contrast. Finally, the circuit successfully produces Y_{sub}, demonstrating the elimination of the overshadowing effect.

Future work will enhance the circuit optimization metric $\mathcal{M}$ by incorporating Y_{sub}'s absolute logit alongside the logit difference with Y_{dom} for more effective guidance. Developing a comprehensive evaluation framework for circuit-based recovery is also crucial. These steps will evolve PHANTOMCIRCUIT into an integrated platform for efficient analysis and robust optimization of knowledge overshadowing.

5 Conclusion

This paper investigates hallucinations in LLMs caused by knowledge overshadowing, and introduces PHANTOMCIRCUIT, a novel knowledge circuit-based analysis framework. PHANTOMCIRCUIT first analyzes the training dynamics of overshadowing, finding that dominant knowledge popularity, model size, and dataset size critically shape the onset, duration, and recovery of overshadowing. Apart from that, the persistent overshadowing in pretrained models stems from inadequately optimized subordinate knowledge loss. By analyzing knowledge circuits, we find that changes in critical attention heads' focus on subordinate knowledge directly correlate with the recovery or onset of overshadowing. Finally, optimizing these knowledge circuits presents a promising strategy for mitigating knowledge overshadowing.

Limitations

Despite the insights provided by PHANTOMCIRCUIT, this study has several limitations that open avenues for future research:

1. The dynamic analysis of knowledge circuits throughout training is computationally intensive, potentially hindering scalability to very large models or extensive training. We aim to develop more computationally efficient techniques for approximating circuit evolution, such as checkpoint-based analysis or lightweight probing.

2. This study concentrates on a specific type of knowledge overshadowing, leaving more complex or subtle interference patterns unaddressed. Future work will broaden PHANTOMCIRCUIT's scope to investigate a wider range of overshadowing phenomena, including those in multi-hop reasoning.

3. Future efforts will focus on evolving our instance-specific circuit optimizations into a generalized mitigation toolkit, supported by a comprehensive evaluation framework. Key improvements will target the precision of automated overshadowed knowledge identification and the broader efficacy of circuit-based interventions. Ultimately, we aim to develop PHANTOMCIRCUIT as a robust platform for both in-depth analysis and effective, generalizable overshadowing mitigation.

References

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):1–15.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Neeloy Chakraborty, Melkior Ornik, and Katherine Driggs-Campbell. 2025. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *ACM Computing Surveys*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, and 1 others. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, and 1 others. 2025. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.

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.

Yunkai Dang, Mengxi Gao, Yibo Yan, Xin Zou, Yanggan Gu, Aiwei Liu, and Xuming Hu. 2024. Exploring response uncertainty in mllms: An empirical evaluation under misleading scenarios. *arXiv preprint arXiv:2411.02708*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Ji-Rong Wen, and Zhicheng Dou. 2025. Understand what llm needs: Dual preference alignment for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, pages 4206–4225.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, and 1 others. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and

1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Kaichen Huang, Jiahao Huo, Yibo Yan, Kun Wang, Yutao Yue, and Xuming Hu. 2024a. Miner: Mining the underlying pattern of modality-specific neurons in multimodal large language models. *arXiv preprint arXiv:2410.04819*.

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

Xinting Huang, Madhur Panwar, Navin Goyal, and Michael Hahn. 2024b. Inversionview: A general-purpose method for reading information from neural activations. *arXiv preprint arXiv:2405.17653*.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *arXiv preprint arXiv:2406.11193*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Callie Y Kim, Christine P Lee, and Bilge Mutlu. 2024. Understanding large-language model (llm)-powered human-robot interaction. In *Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction*, pages 371–380.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*.

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. 2024. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *arXiv preprint arXiv:2404.11584*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

nostalgebraist. 2020. Interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun Chen. 2025. How do llms acquire new knowledge? a knowledge circuits perspective on continual pre-training. *arXiv preprint arXiv:2502.11196*.

Zhixuan Pan, Shaowen Wang, and Jian Li. 2025. Understanding llm behaviors via compression: Data generation, knowledge acquisition and scaling laws. *arXiv preprint arXiv:2504.09597*.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. 2024. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Jiamin Su, Yibo Yan, Fangteng Fu, Han Zhang, Jingheng Ye, Xiang Liu, Jiahao Huo, Huiyu Zhou, and Xuming Hu. 2025. Essayjudge: A multi-granular benchmark for assessing automated essay scoring capabilities of multimodal large language models. *arXiv preprint arXiv:2502.11916*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Aaquib Syed, Can Rager, and Arthur Conmy. 2023. Attribution patching outperforms automated circuit discovery. *arXiv preprint arXiv:2310.10348*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7601–7614.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, and 1 others. 2025. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.

Shuhang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, Linghao Zhang, Shikang Wang, and 1 others. 2025. Rtv-bench: Benchmarking mllm continuous perception, understanding and reasoning through real-time video. *arXiv preprint arXiv:2505.02064*.

Yibo Yan and Joey Lee. 2024. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4163–4167.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2024a. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. *arXiv preprint arXiv:2412.11936*.

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, and 1 others. 2024b. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*.

Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. 2025a. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*.

Yibo Yan, Shen Wang, Jiahao Huo, Philip S Yu, Xuming Hu, and Qingsong Wen. 2025b. Mathagent: Leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. *arXiv preprint arXiv:2503.18132*.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024c. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*, pages 4006–4017.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *arXiv preprint arXiv:2405.17969*.

Murong Yue. 2025. A survey of large language model agents for question answering. *arXiv preprint arXiv:2503.19213*.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024a. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *arXiv preprint arXiv:2402.09267*.

Yuji Zhang, Sha Li, Jiateng Liu, Pengfei Yu, Yi R Fung, Jing Li, Manling Li, and Heng Ji. 2024b. Knowledge overshadowing causes amalgamated hallucination in large language models. *arXiv preprint arXiv:2407.08039*.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R Fung, Kathleen McKeown, Chengxiang Zhai, Manling Li, and 1 others. 2025a. The law of knowledge overshadowing: Towards understanding, predicting, and preventing llm hallucination. *arXiv preprint arXiv:2502.16143*.

Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, and 1 others. 2025b. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. *ACM Computing Surveys*, 57(8):1–39.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. 2024. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. *arXiv preprint arXiv:2408.09429*.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2024. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*.

Junyi Zhu, Shuochen Liu, Yu Yu, Bo Tang, Yibo Yan, Zhiyu Li, Feiyu Xiong, Tong Xu, and Matthew B Blaschko. 2024. Fastmem: fast memorization of prompt improves context awareness of large language models. *arXiv preprint arXiv:2406.16069*.

Xin Zou, Yizhou Wang, Yibo Yan, Sirui Huang, Kening Zheng, Junkai Chen, Chang Tang, and Xuming Hu. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.

A More Related Work

A.1 Large Language Models

First proposed by Brown et al. (2020), Transformer-based auto-regressive LLMs have demonstrated strong performance across a variety of NLP tasks, including question answering (Yue, 2025), in-context learning (Dong et al., 2022), and analogical reasoning (Webb et al., 2023). Pre-trained on large-scale text corpora, LLMs have acquired extensive real-world knowledge from web sources. As a result, models such as InternLM2.5 (Cai et al., 2024), Qwen2.5 (Team, 2024), and LLaMA3.3 (Grattafiori et al., 2024) have shown excellent performance on world knowledge benchmarks (Suzgun et al., 2022). Therefore, over the past year, LLMs have demonstrated remarkable capabilities in understanding-related tasks across various fields (Kim et al., 2024; Nam et al., 2024; Yan et al., 2024c; Yan and Lee, 2024; Yan et al., 2024b; Dong et al., 2025; Su et al., 2025).

Recently, there has been a growing trend toward enhancing LLMs' reasoning capabilities on complex tasks (Guo et al., 2025; Jaech et al., 2024) by generating long Chain-of-Thoughts (CoTs), with reinforcement learning (RL) emerging as an effective tool to encourage this behavior (Li et al., 2025; Trung et al., 2024). Recently, there have also been efforts to explore collaboration between LLMs to enhance their reasoning abilities (Zhang et al., 2025b; Putta et al., 2024; Masterman et al., 2024; Yan et al., 2025b; Chu et al., 2025).

Despite these advancements, existing LLMs still suffer from factual hallucinations in practice (Pan et al., 2025; Asgari et al., 2025), with knowledge overshadowing identified as a primary contributing factor (Zhang et al., 2024b). While existing interoperability works make great efforts on the mechanism of LLM training and generating (Zhao et al., 2024), most of them solely focus on isolated model versions like GPT2 (Wang et al., 2023) and LLaMA2 (Wendler et al., 2024; Tang et al., 2024).

In this paper, we utilize the Pythia suite (Biderman et al., 2023) to investigate the evolution and underlying mechanisms of knowledge overshadowing across models of varying sizes: 70M, 410M, 1.4B, and 2.8B parameters. Sharing a unified architecture, this model suite eliminates design variability, thereby providing clearer and more reliable insights into the scaling behavior of the knowledge overshadowing phenomenon in LLMs.

B PhantomCircuit Details

B.1 Circuit Construction

Knowledge circuit is as a sparse computational subgraph within the LLMs. The construction of such a circuit involves identifying and retaining the most influential components (nodes, including MLPs and attention heads) and connections (edges) while pruning less critical ones.

We adapted the optimized circuit construction method provided by (Yao et al., 2024). The process begins by representing the LLM as a directed acyclic graph (DAG), $G = (V, E)$, where V encompasses input embeddings, attention heads, MLP layers, and output logits, and E represents the information flow between these components. The goal is to identify a subgraph $C \subseteq G$ that is critical for recognize the key component of a given input prompt, particularly in knowledge overshadowing, is X_{dom} and X_{sub}, the difference between P_{dom} and P_{sub}.

The adapted construction method is similar to edge attribution patching(EAP) (Conmy et al., 2023), which involves:

1. **Paired Inputs:** For a given background X_{bg}, we create two primary input prompts: $P_{dom} = (X_{bg}, X_{dom})$ and $P_{sub} = (X_{bg}, X_{sub})$. We also consider a "corrupted" version of P_{sub}, which could be P_{dom} itself or another prompt designed to elicit Y_{dom}. Let's denote the "clean" input as P_{clean} (typically P_{sub}) and the "corrupted" input as P_{corr} (designed to lead to Y_{dom}).

2. **Activation Difference Calculation:** We run both P_{clean} and P_{corr} through the model. For each node $v \in V$ that is a potential parent in an edge, we record its output activation. The difference in activations between the clean and corrupted runs for a node v_p (parent) is denoted as $\Delta A(v_p) = A_{clean}(v_p) - A_{corr}(v_p)$.

3. **Edge Scoring via Gradient-based Attribution:** To score an edge $e = (v_p, v_c)$ (from parent v_p to child v_c), we focus on how patching the activation from v_p (i.e., using $A_{clean}(v_p)$ instead of $A_{corr}(v_p)$ when P_{corr} is the main input) affects a chosen metric $\mathcal{M}$. This metric $\mathcal{M}$ is designed to measure the model's tendency towards generating Y_{sub} versus Y_{dom} when the input is P_{sub}. A common choice for $\mathcal{M}$ could be the logit difference between Y_{sub}

and Y_{dom} at the final layer, or a metric related to our Relative Overshadowing rate (RO).

The score $S(e)$ for an edge e can be approximated by the product of the activation difference from its parent node and the gradient of the metric $\mathcal{M}$ with respect to the input of its child node, when the child node receives the "clean" activation from the parent while other inputs are "corrupted":

$$S(e) \approx \mathbb{E}_{P_{sub}}\left[\Delta A(v_p) \cdot \frac{\partial \mathcal{M}(Y_{target}|P_{sub})}{\partial A_{input}(v_c)}\right]$$

where Y_{target} is ideally Y_{sub}. The expectation $\mathbb{E}$ is taken over instances of P_{sub} in our evaluation set Z_{sub}. In practice, methods like Integrated Gradients (IG) are often used to refine this attribution by integrating gradients along a path from a baseline (corrupted) input to the actual (clean) input.

4. **Circuit Pruning:** Based on the calculated scores $S(e)$, edges with scores below a certain threshold τ, or alternatively, edges outside the top-N highest scores, are pruned from the graph G. The remaining nodes and edges form the knowledge circuit C_{sub}.

$$C_{sub} = (V_{sub}, E_{sub})$$

where $E_{sub} = \{e \in E \mid |S(e)| \geq \tau\}$ (or top-N criterion) and V_{sub} consists of nodes connected by edges in E_{sub}.

This constructed circuit C_{sub} is then analyzed to understand how dominant knowledge K_{dom} might overshadow K_{sub} by examining the attentional features and information flow within it, especially when processing P_{sub}.

B.2 Automated Component Identification for Recovery

Identifying the Overshadowed Component X^*_{sub}. A critical precursor to effective circuit-based recovery is the precise identification of the specific component X^*_{sub} within the subordinate prompt P_{sub} that is being overshadowed. This is achieved by adapting the Relative Pointwise Mutual Information (R-PMI) based methodology from (Zhang et al., 2025a, 2024b). The process involves:

Iteratively generating contrastive prompts P'_{sub} by deleting each candidate token X'_{sub} (a potential overshadowed component) from the original P_{sub}.

For each pair (P_{sub}, P'_{sub}), calculating the R-PMI for tokens y_i in the intersection of their top-k next-token candidate sets, $V_{top}(P_{sub}) \cap V_{top}(P'_{sub})$, using

$$\begin{aligned} R\text{-}PMI(y_i; P_{sub}, P'_{sub}) = \; & \log P(y_i|P_{sub}) \\ & - \log P(y_i|P'_{sub}). \end{aligned}$$

Summing only the negative R-PMI values to obtain

$$S_{R\text{-}PMI^-}(P_{sub}, P'_{sub}) = \sum \min(R\text{-}PMI(y_i), 0).$$

The X'_{sub} that yields the minimum (most negative) $S_{R\text{-}PMI^-}$ is identified as the primary overshadowed component, X^*_{sub}. This selection is based on the rationale that removing the true X^*_{sub} most strongly exposes the model's bias towards outputs favored by the dominant knowledge pattern.

Identifying Target Subordinate Output Y_{sub}. The intended subordinate output Y_{sub} is identified by assessing which token from $V_{top}(P_{sub})$ (the top-k candidates for the original prompt $P_{sub} = (X_{bg}, X_{sub})$) exhibits the most significant improvement when the overshadowing influence of background knowledge (X_{bg}) or other non-subordinate components is mitigated. Specifically, we generate contrastive prompts P'_{sub} by masking or altering components of X_{bg} (or other identified non-subordinate elements that contribute to the $X_{bg} \leftrightarrow Y_{dom}$ association) within P_{sub}. Y_{sub} is then the token $y_i \in V_{top}(P_{sub})$ that shows the most substantial rank improvement (or largest increase in log probability) in these modified prompts P'_{sub} compared to its rank in the original P_{sub}. This rank elevation signifies the "unmasking" of the true subordinate answer as the dominant, overshadowing associations are weakened.

Identifying Dominant Output Y_{dom}. The dominant output Y_{dom} is identified as the token that maintains the highest average rank across all contrastive prompts P'_{sub} generated by deleting different candidate tokens X'_{sub} from P_{sub}. This token represents the model's most consistent, default output tendency when specific subordinate cues are variably weakened, likely reflecting the pervasive influence of dominant knowledge associated with the background X_{bg}.

With X_{bg} (background knowledge), X_{sub} (identified subordinate component), the expected Y_{sub}, and the interfering Y_{dom} established, we prepare the paired inputs required for knowledge circuit construction. The **clean input** is the original subordinate prompt $P_{sub} = (X_{bg}, X_{sub})$, for which the desired output is Y_{sub}. To create the **corrupt input** P_{dom}, which is designed to elicit the overshadowing effect and output Y_{dom}, we maintain the background knowledge X_{bg} but replace the subordinate component X_{sub} with a generic placeholder token, such as 'something'. Thus, $P_{dom} = (X_{bg}, \text{"something"})$. This specific formulation of P_{dom} ensures that while the input structure is similar to P_{sub}, the absence of X_{sub} allows the strong $X_{bg} \leftrightarrow Y_{dom}$ association to dominate, leading to the incorrect prediction Y_{dom}. These paired inputs, P_{sub} and P_{dom}, then serve as the foundation for the activation difference calculations in our circuit analysis.

Some more circuit-based overshadowing recovery cases are shown in Table 1.

C Dataset Details

C.1 Detailed Synthetic Dataset Construction

The synthetic dataset was constructed through the following steps to ensure controlled conditions for analyzing knowledge overshadowing dynamics:

Fixing text lengths. For all generated data instances, consistent token lengths are maintained. The background knowledge (X_{bg}) was set to a length of 4 tokens. All other core components, namely the dominant knowledge entity (X_{dom}), subordinate knowledge entity (X_{sub}), dominant output (Y_{dom}), and subordinate output (Y_{sub}), are each set to a length of 1 token.

Dataset generation for specific D and P Combinations. For each defined combination of total dataset size (D) and knowledge popularity (P), the dataset was built as follows: The dataset comprises multiple distinct groups of knowledge instances. Each group consists of P+1 knowledge prompts: a set of P dominant knowledge prompts $\{P^1_{dom}, P^2_{dom}, \ldots, P^P_{dom}\}$ and one subordinate knowledge prompt P_{sub}. Within each group:

- For the P dominant prompts, the actual dominant knowledge entities $\{X^1_{dom}, X^2_{dom}, \ldots, X^P_{dom}\}$ are all unique. However, they all share the *same* background knowledge component (X_{bg}) and are associated with the *same* dominant output (Y_{dom_g}). Thus, each $P^i_{dom} = (X_{bg}, X^i_{dom})$ is paired with Y^g_{dom}.
- The single subordinate prompt $P_{sub} = (X_{bg}, X_{sub})$ uses the *same* background knowledge X_{bg} as the dominant prompts in that group. However, its subordinate knowledge entity X_{sub} is distinct from all X^i_{dom} entities in that group, and its corresponding output Y_{sub} is distinct from any Y_{dom} in group.

This structure creates a group: $\{(P^1_{dom}, Y^1_{dom}), \ldots, ((P^P_{dom}, Y^P_{dom}), (P_{sub}, Y_{sub})\}$.

Multiple such groups are generated. All tokens for $X_{bg}, X^i_{dom}, X_{sub}, Y_{dom}, Y_{sub}$ within each group, and across different groups, are randomly sampled from the Pythia tokenizer vocabulary, ensuring no overlap between the core entities of different groups. This process was repeated until the total number of tokens in the dataset reached the target size D.

Cases illustration. We illustrate some groups for P=5 dataset in Table 2. We directly show token id.

C.2 Finetuning dataset

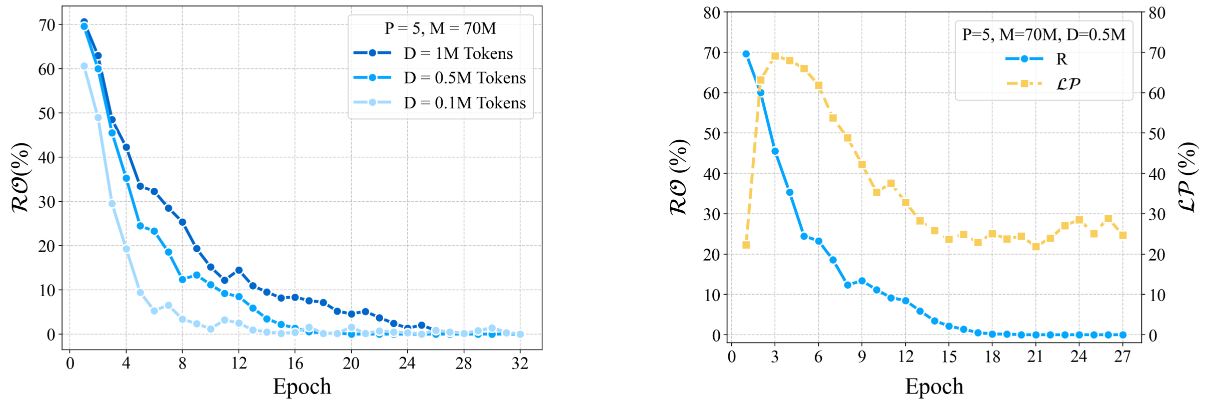
For the finetuning dataset, we utilized the Qwen-Long API to generate instances of virtual knowledge. This generated data subsequently underwent manual review to identify and remove any instances that are overly repetitive or semantically too similar, ensuring a degree of diversity, resulted in D = 1M.

A key distinction from the synthetic dataset construction is that we did not strictly control token lengths for each component in this dataset. Instead of randomly sampled token IDs, the finetuning dataset consists of actual linguistic statements that, while syntactically and semantically coherent, represent virtual (i.e., fabricated but plausible) knowledge. The underlying pattern of dominant and subordinate knowledge construction, however, mirrors that of the synthetic dataset.

As an example of this dataset, we set the knowledge popularity P=5. Some illustrative cases from the dataset are shown in Table 3.

D More Dynamics Analysis on Finetuning Dataset

In addition to validating the efficacy of our circuit-based overshadowing recovery method, the finetuning dataset serves a dual purpose. We also leverage it to empirically verify our conclusions regarding the training dynamics of knowledge overshadowing, specifically concerning the impact of Dataset Size (D). Consistent with our dynamic analysis findings, we investigate whether a larger D indeed



(a) The $\mathcal{RO}$ during training phase of the different finetuning dataset size (D).

(b) The co-evolution of $\mathcal{LP}$ and $\mathcal{RO}$ during training phase on finetuning dataset.

Figure 6: The dynamics analysis of knowledge overshadowing in finetuning dataset.

correlates with a slower recovery rate from knowledge overshadowing and a prolonged duration of the hallucination effect. To this end, we conduct experiments on the finetuning dataset by fixing Knowledge Popularity at P=5 and Model Size at M=70M, while varying D across values of 0.1M, 0.5M, and 1M tokens. The results, as depicted in Figure 6a, corroborate this relationship. Furthermore, under the specific configuration of P=5, M=70M, and D=0.5M on the finetuning dataset, we re-examine the interplay between the loss proportion of subordinate knowledge ($\mathcal{LP}$) and the relative overshadowing rate ($\mathcal{RO}$). As shown in Figure 6b, the observations again support the hypothesis that insufficient optimization of subordinate knowledge contributes to the persistence of knowledge overshadowing.

It is noteworthy that distinct behaviors are observed when comparing the finetuning dataset to the synthetic dataset. Firstly, the recovery from overshadowing on the finetuning dataset is generally slower than on the synthetic dataset for same D. This can be attributed to the richer semantic relationships and greater complexity inherent in the natural language of the finetuning data, which presents a more challenging learning task.

Secondly, we observe that the finetuning dataset exhibits a minimal or absent onset phase for knowledge overshadowing, where $\mathcal{RO}$ typically rise. This is because finetuning commences from a pre-trained model, which has already moved beyond the initial epochs of chaotic, random predictions. Consequently, the model can very rapidly generalize strong association patterns present in the finetuning data. Moreover, the diverse and varied forms of data within the finetuning set may act akin to a beneficial noise signal, prompting the model to pay closer attention to distinguishing features and differences. This inherent data diversity can help preemptively mitigate or even eliminate the early onset stage of knowledge overshadowing that might otherwise be observed.

Table 1: Circuit-based overshadowing recovery cases

<table>
<tr><th>Case</th><th>P_{sub} with $\{X_{sub}\}$</th><th>$\mathcal{M}$ indicator (logits difference)</th><th>Y_{dom} & Y_{sub}</th><th>Full Model Top 5 Prediction</th><th>Circuit Top 5 Prediction</th></tr>
<tr><td>Case 1</td><td>Analysis of the Chrono-Filter device efficiency for temporal sorting shows outcome {filtration overload}</td><td>Original model:-1.283 & Circuit: 0.764</td><td>|Time| & |Tem|</td><td>Rank 0: Logit: 16.18 Prob: 32.18% Token: |Tem|
Rank 1: Logit: 15.42 Prob: 14.98% Token: |Time|
Rank 2: Logit: 14.21 Prob: 4.47% Token: |Sp|
Rank 3: Logit: 14.05 Prob: 3.83% Token: |T|
Rank 4: Logit: 13.84 Prob: 3.10% Token: |Custom|</td><td>Rank 0: Logit: 21.03 Prob: 73.70% Token: |Time|
Rank 1: Logit: 19.75 Prob: 20.42% Token: |Tem|
Rank 2: Logit: 17.34 Prob: 1.84% Token: |Filter|
Rank 3: Logit: 16.40 Prob: 0.72% Token: |E|
Rank 4: Logit: 16.29 Prob: 0.64% Token: |Sp|</td></tr>
<tr><td>Case 2</td><td>Constructing psionic wave emitters necessitates precise tuning involving specialized harmonic {feedback loop}</td><td>Original model: -0.745& Circuit: 2.490</td><td>|Wave| & |E|</td><td>Rank 0: Logit: 17.64 Prob: 37.18% Token: |Wave|
Rank 1: Logit: 16.89 Prob: 17.65% Token: |E|
Rank 2: Logit: 15.35 Prob: 3.76% Token: |Ps|
Rank 3: Logit: 15.29 Prob: 3.57% Token: |St|
Rank 4: Logit: 15.15 Prob: 3.09% Token: |emitter|</td><td>Rank 0: Logit: 17.96 Prob: 46.34% Token: |E|
Rank 1: Logit: 16.47 Prob: 10.39% Token: |Ps|
Rank 2: Logit: 15.47 Prob: 3.84% Token: |Wave|
Rank 3: Logit: 15.31 Prob: 3.28% Token: |Energy|
Rank 4: Logit: 14.89 Prob: 2.15% Token: |emitter|</td></tr>
<tr><td>Case 3</td><td>Analyzing Ectoplasmic Conduit energy transfer efficiency through degrading {structure reinforcement reveals}</td><td>Original model: -3.019 & Circuit: 0.523</td><td>|Transfer| & |Emit|</td><td>Rank 0: Logit: 49.08 Prob: 49.26% Token: |Transfer|
Rank 1: Logit: 48.68 Prob: 33.09% Token: |St|
Rank 2: Logit: 46.66 Prob: 4.38% Token: |Ada|
Rank 3: Logit: 46.37 Prob: 3.28% Token: |Flow|
Rank 4: Logit: 46.06 Prob: 2.41% Token: |Emit|</td><td>Rank 0: Logit: 44.54 Prob: 47.18% Token: |Emit|
Rank 1: Logit: 44.01 Prob: 27.96% Token: |Transfer|
Rank 2: Logit: 42.38 Prob: 5.48% Token: |St|
Rank 3: Logit: 42.24 Prob: 4.75% Token: |Energy|
Rank 4: Logit: 41.64 Prob: 2.61% Token: |Mi|</td></tr>
<tr><td>Case 4</td><td>Shard Relic residual energy output response to sudden energy {conduit field spikes} shows</td><td>Original model: -2.623 & Circuit: 3.202</td><td>|Output| & |St|</td><td>Rank 0: Logit: 40.96 Prob: 91.94% Token: |Output|
Rank 1: Logit: 38.34 Prob: 6.67% Token: |St|
Rank 2: Logit: 35.87 Prob: 0.56% Token: |Fl|
Rank 3: Logit: 35.82 Prob: 0.54% Token: |Trans|
Rank 4: Logit: 33.51 Prob: 0.05% Token: |Un|</td><td>Rank 0: Logit: 34.61 Prob: 51.02% Token: |St|
Rank 1: Logit: 34.24 Prob: 35.19% Token: |Field|
Rank 2: Logit: 31.41 Prob: 2.08% Token: |Output|
Rank 3: Logit: 30.97 Prob: 1.34% Token: |Har|
Rank 4: Logit: 30.84 Prob: 1.18% Token: |Fl|</td></tr>
</table>

Table 2: Illustrative examples from the synthetic dataset (P=5). Each data entry is a row, with fine lines separating entries within a group. Token IDs are shown.

Group	X_{bg}	X_{dom}	Y_{dom}	X_{sub}	Y_{sub}
Group 1	[10030, 16936, 1050, 10565]	10279	20730		
	[10030, 16936, 1050, 10565]	24327	20730		
	[10030, 16936, 1050, 10565]	4619	20730		
	[10030, 16936, 1050, 10565]	5137	20730		
	[10030, 16936, 1050, 10565]	785	20730		
	[10030, 16936, 1050, 10565]			18941	3519
Group 2	[17026, 8837, 3802, 28741]	2496	1077		
	[17026, 8837, 3802, 28741]	3530	1077		
	[17026, 8837, 3802, 28741]	11948	1077		
	[17026, 8837, 3802, 28741]	2028	1077		
	[17026, 8837, 3802, 28741]	9389	1077		
	[17026, 8837, 3802, 28741]			25814	5374
Group 3	[18131, 14501, 21161, 311]	4706	7790		
	[18131, 14501, 21161, 311]	778	7790		
	[18131, 14501, 21161, 311]	18762	7790		
	[18131, 14501, 21161, 311]	28591	7790		
	[18131, 14501, 21161, 311]	28981	7790		
	[18131, 14501, 21161, 311]			8447	6129

Table 3: Illustrative examples from the Finetuning Dataset (P=5). Each data entry is a row, with fine lines separating entries within a group.

Group	X_{bg}	X_{dom}	Y_{dom}	X_{sub}	Y_{sub}
Group 1	Analysis of the Starfire Crystal Engine reveals primary energy	output peak resonance	Pure Nova		
	Analysis of the Starfire Crystal Engine reveals primary energy	output idle cycle	Pure Nova		
	Analysis of the Starfire Crystal Engine reveals primary energy	output phase synchronicity	Pure Nova		
	Analysis of the Starfire Crystal Engine reveals primary energy	output null gravity	Pure Nova		
	Analysis of the Starfire Crystal Engine reveals primary energy	output core temperature	Pure Nova		
	Analysis of the Starfire Crystal Engine reveals primary energy			output cold ignition	Quantum Flux
Group 2	Observation of the Chronos Orb's temporal distortion signature	indicates full moonlight	Stable Chrono-displacement		
	Observation of the Chronos Orb's temporal distortion signature	indicates near Leyline	Stable Chrono-displacement		
	Observation of the Chronos Orb's temporal distortion signature	indicates while attuned	Stable Chrono-displacement		
	Observation of the Chronos Orb's temporal distortion signature	indicates during solstice	Stable Chrono-displacement		
	Observation of the Chronos Orb's temporal distortion signature	indicates etherium interaction	Stable Chrono-displacement		
	Observation of the Chronos Orb's temporal distortion signature			indicates sudden impact	Temporal Anomaly
Group 3	Study of the Xylosian biology reveals primary nutrient absorption	method consuming light	Molecular Assimilation		
	Study of the Xylosian biology reveals primary nutrient absorption	method after sonic bath	Molecular Assimilation		
	Study of the Xylosian biology reveals primary nutrient absorption	method during digestion	Molecular Assimilation		
	Study of the Xylosian biology reveals primary nutrient absorption	method high pressure	Molecular Assimilation		
	Study of the Xylosian biology reveals primary nutrient absorption	method thermal vent	Molecular Assimilation		
	Study of the Xylosian biology reveals primary nutrient absorption			method xenoflora consumption	Crystalline Excretion