Budgeted Causal Selection for Neuron-Level Knowledge Attribution in Large Language Models

Anonymous ACL submission

Abstract

Existing neuron attribution methods face a trade-off between causal reliability and computational efficiency:dynamic intervention-based methods provide stronger causal evidence but are computationally expensive, while static methods are efficient yet largely correlational. To address this issue, we formulate neuron-level knowledge attribution as a Budgeted Causal Selection problem and propose SDS, a static–dynamic selection framework for efficient causal neuron attribution. SDS first applies high-recall static filtering to reduce the search space, and then performs dynamic causal interventions within the candidate set to identify causally important neurons. Experiments on GPT2-large and Llama-7B across multiple factual knowledge tasks show that SDS consistently outperforms six representative baselines under identical intervention budgets. Further analysis suggests that factual knowledge in large language models is sparsely and structurally distributed, where perturbing as little as 1% of neurons can significantly alter model predictions.

1 Introduction

Large Language Models (LLMs) based on the Transformer architecture (Vaswani et al., 2017; Brown et al., 2020) have demonstrated remarkable capabilities in storing and retrieving factual knowledge in recent years (Chowdhery et al., 2023) , and have been widely applied to question answering (Achiam et al., 2023) and reasoning tasks (Wei et al., 2022). As model scale continues to grow, understanding how LLMs internally represent and retrieve knowledge (Roberts et al., 2020) has gradually become a central issue in research on model reliability and interpretability (Jiang et al., 2020; Rajani et al., 2019; Sharkey et al., 2025). Existing studies primarily conduct knowledge attribution analysis at the neuron level (Dai et al., 2022; Chen et al., 2025), and can be divided intothe static method and the dynamic method, as shown in Figure 1:Static attribution methods compute importance using model activations or structural parameters (Sundararajan et al., 2017; Selvaraju et al., 2016). Although they offer excellent scalability, they struggle to guarantee the faithfulness and causal accuracy of explanations due to the lack of interventions (Jacovi and Goldberg, 2020; Hooker et al., 2019); In contrast, dynamic attribution methods measure the causal contribution of neurons through interventions such as ablation or replacement (Meng et al., 2022; Vig et al., 2020; Fartale et al., 2025). While they achieve higher attribution accuracy, their computational cost grows explosively with the size of the neuron space (Hase et al., 2023; Yu and Ananiadou, 2024).

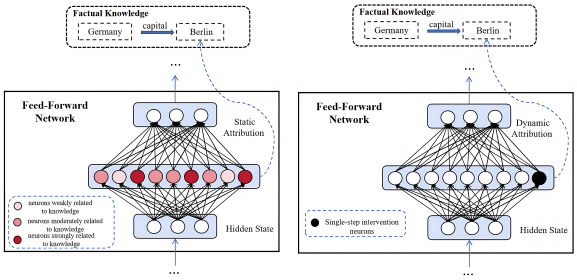

Figure 1: The static method (left) computes importance scores for all neurons simultaneously based on a single forward/backward pass. In contrast, the dynamic method (right) assesses causal contributions through targeted interventions, typically evaluating a specific subset of neurons per iteration.

In this paper, we rethink neuron attribution as a budgeted causal selection problem: maximizing causal explanatory capability within a finite intervention budget N. Finding the optimal subset of neurons is a combinatorial optimization challenge that is computationally NP-hard due to the explosive growth of the neuron space.

To address this challenge, we propose static–dynamic selection method(SDS), a two-

stage framework that balances efficiency and causal precision. SDS first employs lightweight static attribution metrics to narrow the candidate set and construct a high-recall candidate set. It then performs dynamic causal interventions only within this reduced subspace to identify neurons with the strongest causal effects on factual predictions. By transforming global combinatorial search into localized causal optimization, SDS effectively combines the scalability of static attribution with the reliability of dynamic intervention.

On six factual knowledge benchmarks, we compare SDS with six existing baselines under a budget-constrained neuron selection setting, where the goal is to identify at most n neurons for causal intervention and attribution. Results show that, in low-budget settings with limited neuron ablation, SDS consistently outperforms all baselines across three evaluation metrics, indicating its stronger ability to approximate informative causal neuron subsets under resource constraints. We further evaluate SDS across different intervention budgets n and observe stable performance gains, demonstrating robustness under varying constraint levels. In addition, analysis over the expansion ratio α shows that moderate candidate expansion achieves the best trade-off between computational cost and causal recall. Finally, we find that neurons selected by SDS exhibit consistent clustering patterns in deeper FFN layers and higher-level attention modules, suggesting a hierarchical organization of knowledge representations in Transformer models.

In summary, our main contributions are as follows:

a) We propose a static–dynamic selection neuron selection framework that formulates neuron attribution as a budget-constrained combinatorial optimization problem, enabling efficient and reliable causal attribution at scale;

b) We analyze the effects of candidate set size and intervention strength, showing stable performance under varying resource constraints;

c) We identify consistent layer-wise distribution patterns of attention and FFN neurons across tasks, suggesting an implicit prior structure in Transformer knowledge representations.

2 Related Work

2.1 Static Attribution

Static attribution does not require any intervention in the model's inference process, making it computationally inexpensive and easily scalable. Common approaches primarily include activation-based (Geva et al., 2021; Dai et al., 2022; Geva et al., 2022) and parameter-based (Yu and Ananiadou, 2024) attribution methods, which infer the potential contribution of neurons or weights to the model output by observing the response patterns of internal features. For example, Geva et al. (2021) found that the FFNs of Transformers can be viewed as key–value memory, where certain neurons exhibit high activation when processing specific factual knowledge. Similarly, Dai et al. (2022) pointed out that "knowledge neurons" exist in pretrained Transformers, whose activation patterns reveal the model's memory of particular facts.

Static attribution methods can be used to rapidly generate candidate neuron sets and guide subsequent analyses (Voita et al., 2019a; Geva et al., 2021; Dai et al., 2022). They are also widely applied to understand the functional division between attention heads and FFN layers in knowledge storage (Geva et al., 2022; Voita et al., 2019b; Yu and Ananiadou, 2024) . However, due to the lack of intervention in the model's decision process, the causal interpretability of static methods is limited; highly activated neurons are not necessarily responsible for the output (Hase et al., 2023; Jacovi and Goldberg, 2020; Vig et al., 2020) . Therefore, relying solely on static attribution may not accurately reflect the true role of neurons during actual reasoning.

2.2 Dynamic Attribution

Compared with the limitations of static attribution, dynamic attribution methods measure the true causal effect by directly intervening on model activations, replacing internal states, or ablating specific neurons (Neuberg, 2003). Representative approaches include causal mediation analysis (Vig et al., 2020) and activation patching as a causal probe (Wang et al., 2022). In the context of knowledge localization, methods such as Knowledge Neurons (Dai et al., 2022) and the ROME series (Meng et al., 2022) demonstrate that interventions on a model's internal structure can effectively locate and modify factual knowledge. Meanwhile, research on mechanistic interpretability further suggests that Transformers contain decomposable causal pathways and feature circuits (Elhage et al., 2021), providing additional theoretical support for dynamic methods in understanding knowledge storage within models (Nanda et al.,

2023; Olsson et al., 2022).

Although dynamic ablation and activation interventions offer stronger causal guarantees, their main drawback is the high computational cost (Meng et al., 2022; Grosse et al., 2023). Consequently, relying solely on dynamic methods is impractical for large-scale knowledge localization tasks, and reducing attribution cost while maintaining causal reliability has become a central research focus in recent years (Hase et al., 2023; Yao et al., 2023).

3 Problem Formulation

This section formalizes neuron attribution as a constrained optimization problem. We first define neuron-level representations (Section 3.1) and then formulate the budgeted causal selection problem (Section 3.2).

3.1 Neuron-Level Representations in Transformers

A Transformer consists of stacked self-attention modules and FFNs, where the outputs of each submodule are accumulated through residual connections to form the final hidden representations (Xiong et al., 2020). In this paper, we define a neuron as a basic computational unit that contributes an additive vector to the residual stream during the forward pass (Geva et al., 2021). Within FFNs, neurons correspond to nonlinearly activated units together with their output vectors. Specifically, for the FFN in layer l, let m_k^l denote the activation of the k-th neuron, and let v_k^l denote its corresponding output vector. The contribution of neuron u to the residual stream can be written as: $u = m_k^l v_k^l$. This injects a vector with a fixed direction and input-dependent magnitude into the residual stream.

Within attention modules, fine-grained computational units can be interpreted as value–output vectors. Specifically, each value vector contributes additively to the residual stream after attention weighting and output projection, where the contribution is determined by the attention distribution (Yu and Ananiadou, 2024).

This additive structure enables the effect of individual neurons to be independently characterized in the representation space, thereby providing a foundation for neuron-level analysis.

3.2 The Budgeted Causal Selection Problem

Let the input query be H, the model be f_θ, and the target knowledge token be y^*. The set of all neurons is $U = \{u_1, u_2, \ldots, u_K\}$. We define the causal importance of neuron u_i as the change in prediction probability when u_i is suppressed (zero-out):

$$G(u_i) = p(y^*|H) - p(y^*|H; u_i = 0) \quad (1)$$

In practice, we aim to identify a subset of neurons S that captures the maximum causal impact. However, exhaustive causal intervention is computationally prohibitive. Therefore, we define the budgeted causal selection problem: Given a strict intervention budget N (where $N \ll K$), find an optimal subset S^* such that:

$$S^* = \arg \max_{S \subseteq U, |S| \leq N} \sum_{u \in S} G(u) \quad (2)$$

Identifying S^* requires $O(K)$ forward passes if done naively. Our goal is to design an approximation algorithm that solves this with $O(N)$ or $O(\alpha N)$ complexity.

4 Static–Dynamic Selection for Budgeted Causal Attribution

To solve the budgeted causal selection problem defined in Eq.(2), we propose the Static–Dynamic Selection Framework(SDS). Rather than performing prohibitively expensive global causal search, SDS formulates neuron attribution as a "proposal-then-verify" process: static attribution serves as a lightweight proposal mechanism to efficiently narrow the candidate set, while dynamic causal intervention verifies and refines the final neuron subset. In this way, SDS efficiently approximates the optimal causal subset S^* under a fixed intervention budget.

Phase 1: Static Candidate Space Reduction. Assume a given input sequence:

$$H = [t_1, t_2, \ldots, t_T] \quad (3)$$

After l layers of computation by the Transformer model, the final hidden representation h_T^l is obtained at the target position T. This representation is projected into the vocabulary space via the unembedding matrix E_u to obtain the logit for each candidate token y^* before the softmax, denoted as the Before-Softmax Value (BS-Value) (Brown et al., 2020; Radford et al., 2019):

$$bs_{y^*}(H) = e_{y^*}^\top h_T^l \quad (4)$$

Where e_{y^*} is the row vector in E_u corresponding to token y^*. The model's predicted probability for

the target token is jointly determined by all token BS-Values after normalization "softmax".

Since the outputs of the attention modules and FFNs in a Transformer are accumulated layer-by-layer through residual connections via vector addition (Elhage et al., 2021), the final representation can be viewed as the result of a linear superposition of multiple fine-grained computational units (Nanda et al., 2023; Geva et al., 2021, 2022).

Based on this perspective of vector additivity, the final hidden representation can be abstracted into a bipartite form consisting of the target neuron vector and the contributions from other computational paths:

$$h_T^l = u + x \tag{5}$$

Where u represents the vector contribution corresponding to the neuron under analysis, and x aggregates the synthetic representation formed by the remaining neurons and residual connections. Under this representation, the "logit increment" of this neuron for the target token y^* is:

$$\Delta bs_{y^*} = bs_{y^*}(u + x) - bs_{y^*}(x) = e_{y^*}^\top u \tag{6}$$

Eq.(6) demonstrates that at the BS-Value level, the impact of a single neuron vector on the output logit possesses strict linear additivity, allowing for independent analysis of a neuron's effect without altering the model's forward propagation path. Furthermore, the static importance of a neuron is defined as the increment in the log-prediction probability of the target token before and after adding its vector to the context (Yu and Ananiadou, 2024):

$$F_{\text{static}}(v) = \log p(y^*|u + x) - \log p(y^*|x) \tag{7}$$

This static score can be obtained relying solely on forward computation, thus offering high computational efficiency and suitability for rapid filtering across the entire neuron space. However, it should be noted that F_{static} is essentially still a correlation-based attribution metric. A high static score only indicates that a certain neuron is highly correlated with the prediction result and does not guarantee its causal necessity for that prediction.

The set of static scores for all neurons is:

$$F_{\text{static}} = (F_{\text{static}}(u_1), \ldots, F_{\text{static}}(u_K)) \tag{8}$$

In the first phase, the top αn neurons are selected from the total neuron pool based on their static scores to form the candidate set:

$$\mathcal{C} = \text{Top}\alpha n(F_{\text{static}}) \tag{9}$$

where $n(n = N)$ is the number of important neurons ultimately targeted for localization, and $\alpha > 1$ is the expansion ratio. Extensive research indicates that knowledge storage within language models exhibits significant sparsity and modularity; that is, a single piece of knowledge is often dominated by a very small number of neurons. Therefore, with a reasonable value of α, the true $\text{Top}n$ causal neurons are included in the candidate set $\mathcal{C}$ with extremely high probability, enabling the first phase to achieve recall at a very low cost.

Phase 2: Dynamic Causal Verification. In the first stage, a candidate set $\mathcal{C}$ is obtained based on static scores. In the second stage, dynamic causal interventions are performed only on the neurons within $\mathcal{C}$ to precisely locate knowledge representations inside the model, thereby completing knowledge attribution.

Specifically, in the second stage, dynamic causal interventions are applied exclusively to the neurons in the candidate set $\mathcal{C}$, and the causal importance $G(u_i)$ is computed for each $u_i \in \mathcal{C}$ as defined in Eq.(1).

This metric measures the degree of change in the prediction probability of the target token after intervening on neuron u_i. Based on this causal importance, the $\text{Top}n$ key neurons are selected:

$$S = \text{Top}n(G(u_i), u_i \in \mathcal{C}) \tag{10}$$

Considering that if an intervention on a particular neuron leads to a significant change in the target prediction probability, it indicates that the neuron has a high causal contribution to the representation of the corresponding knowledge, resulting in a higher importance value. By ranking the importance of the candidate neurons, the core neurons truly involved in knowledge expression can be further filtered. Consequently, the $\text{Top}n$ key neurons S obtained serve as the set of neurons that SDS attributes as having a high causal contribution to factual knowledge learning.

It can be seen that SDS inherits both the high efficiency of static methods and the high accuracy of dynamic methods. In terms of computational complexity, SDS offers a significant advantage over purely dynamic methods. If dynamic interventions were performed on all K neurons, the complexity would be:

$$O(K \cdot \mathcal{C}_{\text{forward}}) \tag{11}$$

In contrast, the total complexity of SDS is:

$$O(\mathcal{C}_{\text{static}}) + O(\alpha n \cdot \mathcal{C}_{\text{forward}}) \tag{12}$$

where $\mathcal{C}_{\text{static}}$ represents the cost of a single static score calculation, and αn is typically only on the order of dozens (for example, when $n = 10$ and $\alpha = 4$, only 40 neurons need to be intervened). Compared to intervening on all neurons, this method reduces complexity by several orders of magnitude while sacrificing almost no causal accuracy of the dynamic approach, thereby making large-scale neuron attribution analysis feasible in practice.

5 Experiments

In this section, we evaluate SDS under a budget-constrained neuron selection setting. Section 5.1 presents comparisons with six representative baselines on causal neuron subset approximation under limited intervention budgets. Section 5.2 analyzes the effects of ablated neuron number n and candidate expansion ratio α, characterizing the behavior of the static–dynamic selection mechanism under different resource constraints. Section 5.3 examines the layer-wise and neuron-level distribution of different types of factual knowledge, revealing structural localization patterns in Transformer representations.

5.1 Comparison of Neuron Attribution Methods

To evaluate SDS under a budget-constrained causal neuron selection setting, we compare the proposed static–dynamic attribution framework with six representative baselines. For each sample, all methods select a fixed-size subset of FFN neurons ($n = 10$), and the quality of the selected causal subset is evaluated using three complementary metrics.

Dataset. We construct question–answer samples from TriviaQA(Joshi et al., 2017), covering six categories of factual knowledge: country, language, capital, color, number, and month. To ensure that the selected samples reflect knowledge already learned by the model—rather than noise introduced by incorrect predictions—we further filter the dataset by retaining only those instances where the correct answer token appears within the model's top-10 predicted tokens.

Models. We conduct experiments on GPT2-large (Radford et al., 2019) and Llama-7B (Touvron et al., 2023). These two models differ significantly in both parameter scale and architectural design, enabling us to evaluate the robustness and consistency of our method across different model sizes, as well as to compare how knowledge is stored in different architectures.

Hyperparameters. Unless otherwise specified, we use an intervention budget of $n = 10$ neurons and set the candidate expansion ratio to $\alpha = 4$, which provides a favorable trade-off between causal attribution quality and computational efficiency.

Attribution Methods. We compare our proposed method with six attribution(Yu and Ananiadou, 2024) baselines, including:
a) SDS (ours)
b) log prob increase: $\log p(y^* \mid x + m_k^l v_k^l) - \log p(y^* \mid x)$
c) log prob increase-random: a randomized variant of b) with stochastic causal intervention.
d) prob increase: $p(y^* \mid x + m_k^l v_k^l) - p(y^* \mid x)$
e) $p(y^* \mid m_k^l v_k^l)$ (Dar et al., 2023)
f) coefficient score: $|m_k^l|$
g) norm: $|v_k^l|$

Evaluation metrics. To evaluate the causal quality of different neuron selection methods under a budget-constrained intervention setting, we intervene on the selected neuron subset and measure the resulting changes in model predictions. Specifically, for each method, we ablate a fixed-size subset of FFN neurons ($n = 10$) by zeroing their activations and recomputing the model outputs. This setting formulates neuron attribution as a constrained causal selection problem, where the objective is to identify the most causally influential neurons under limited intervention budgets.

We adopt three complementary metrics: mean reciprocal rank (MRR) of the correct target token y^*, prediction probability (prob), and log-probability (logp). MRR captures changes in the ranking position of the correct answer, prob reflects model confidence, and logp provides a numerically stable characterization of probability variation on a logarithmic scale. Larger degradation across these metrics indicates that the selected neuron subset exerts stronger causal influence on the model's prediction behavior and knowledge retrieval process.

Experimental Analysis. To evaluate SDS, we compare it with six neuron attribution baselines on Llama-7B and GPT2-large under a budget-constrained intervention setting, using three metrics: MRR, prob, and logp. As shown in Table 1, SDS consistently achieves the largest performance degradation across both models and all metrics, indicating a more accurate approximation of causally critical neuron subsets. Specifically, on Llama-7B,

	Llama-7B			GPT2-large		
	prob	logp	MRR	prob	logp	MRR
o)	0.38	-1.18	0.87	0.032	-3.74	0.27
a)	**0.14**	**-2.48**	**0.54**	**0.011**	**-4.89**	**0.11**
b)	0.21	-2.10	0.64	0.012	-4.64	0.13
c)	0.28	-1.62	0.76	0.018	-4.19	0.18
d)	0.22	-2.02	0.65	0.012	-4.61	0.14
e)	0.23	-1.94	0.68	0.013	-4.64	0.13
f)	0.46	-1.05	0.85	0.033	-3.69	0.28
g)	0.38	-1.18	0.87	0.032	-3.73	0.27

Table 1: Results of attribution methods on two models.

SDS reduces prob from 0.38 to 0.14, logp from -1.18 to -2.48, and MRR from 0.87 to 0.54; on GPT2-large, prob decreases from 0.032 to 0.011, logp from -3.74 to -4.89, and MRR from 0.27 to 0.11. The consistent drop across all three metrics suggests that the neurons identified by SDS are not only highly correlated with predictions but also causally responsible for encoding key knowledge.

Among static methods, those based on probability or log-probability changes (b, d, e) generally outperform metrics relying solely on coefficients or vector norms (f, g), indicating that directly measuring a neuron's impact on the target distribution is a more effective static criterion. However, their performance degradation remains substantially weaker than that of SDS, suggesting that purely static signals struggle to distinguish between highly correlated but non-causal neurons and truly causal ones. Furthermore, log prob increase-random (c), which introduces random dynamic interventions on top of log prob increase (b), performs worse than SDS and even underperforms the targeted static method log prob increase-random (b), demonstrating that dynamic intervention alone does not guarantee effective neuron selection and is highly dependent on the quality of the candidate set. In addition, methods based solely on parameter magnitude (f) or vector norm (g) barely degrade model performance, yielding results close to the original model (o). This suggests that relying only on parameter scale or activation strength is insufficient to capture the causal importance of neurons. Instead, explicitly measuring changes in the prediction distribution under intervention is crucial for distinguishing correlation from causation.

Overall, SDS significantly improves the causal accuracy of neuron attribution while maintaining computational efficiency. This advantage remains consistent across different model scales, demonstrating strong generalizability. Building on these findings, we further conduct extended experiments along two dimensions: (1) analyzing the effect of the number of removed neurons n, and (2) studying the impact of the candidate expansion ratio α.

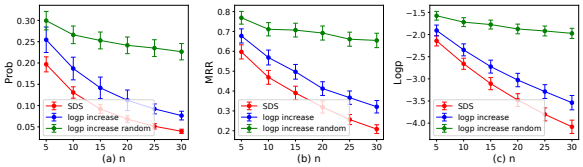

Figure 2: Neuron Ablation Analysis on Llama-7B across Attribution Methods.

5.2 Budget and Candidate Sensitivity Analysis

We analyze the effects of the intervention budget n and the candidate expansion ratio α to understand the behavior of SDS under different resource constraints. This study aims to characterize how the static–dynamic selection process adapts to varying intervention budgets and candidate space configurations in the underlying combinatorial optimization problem.

Effect of Intervention Budget. We analyze the effect of the intervention budget n, i.e., the number of ablated neurons, to study how different methods behave under varying intervention strengths. As shown in Figure 2 (GPT2-large results in Appendix A), all methods exhibit consistent performance degradation in prob, logp, and MRR as n increases, indicating that larger intervention budgets progressively disrupt the model's ability to express target knowledge.

Across all settings, SDS consistently achieves the largest degradation, even under small budgets (e.g., $n = 5, 10$), and maintains its advantage as n increases. In contrast, static probability-based methods produce moderate effects, while random-intervention variants show weak and unstable performance, often comparable to or worse than their static counterparts.

These results indicate that SDS identifies neurons with stronger causal control over both generation probability and ranking behavior. Importantly, the gains are not solely due to restricting the candidate set, but stem from the static–dynamic design, where static filtering reduces the search space and dynamic intervention further removes non-causal yet correlated neurons.

Effect of Candidate Expansion Ratio. We analyze the effect of the candidate expansion ratio

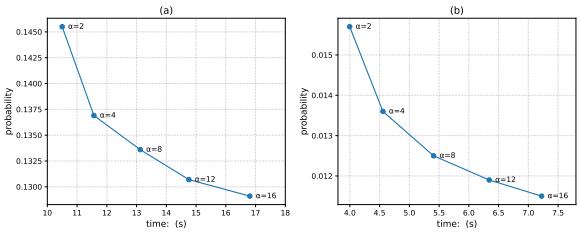

Figure 3: Impact of Expansion Ratio α on Runtime and Prediction Probability.

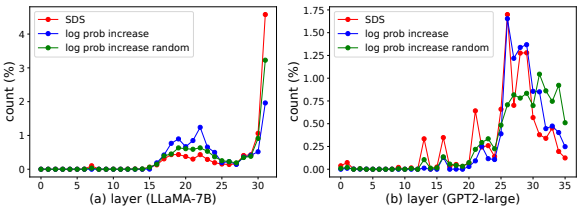

Figure 4: Layer-wise distribution of critical neurons in Llama-7B (left) and GPT2-large (right).

$\alpha \in 2, 4, \ldots, 16$ under the budgeted neuron selection framework. As shown in Figure 2, increasing α consistently improves causal attribution performance, reflected by a monotonic decrease in the target token probability, while introducing higher computational cost due to a larger candidate space for dynamic intervention.

We observe an elbow point at $\alpha = 4$ for both models, where most of the performance gain is achieved, while further increases lead to diminishing returns with additional computational overhead. We therefore use $\alpha = 4$ as the default setting to balance attribution quality and efficiency.

From a complexity perspective, without candidate compression, exhaustive intervention over all FFN neurons ($K \approx 3.5 \times 10^5$ for Llama-7B) would be infeasible. In contrast, SDS reduces the intervention space to αn candidates, significantly reducing the search space for causal verification. This demonstrates that the static filtering stage plays a critical role in making large-scale causal attribution tractable by restricting dynamic interventions to a compact candidate subset.

5.3 Exploration on Different Knowledge Types

We further analyze whether neurons identified by SDS exhibit structural consistency with the internal knowledge organization of large language models. This serves as a structural validation of the proposed budget-constrained causal selection framework.

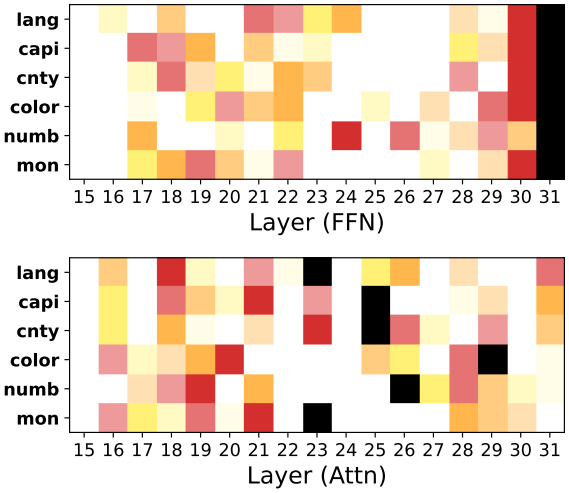

Figure 5: Heatmap of importance scores for the top-10 FFN and Attention layers in Llama-7B.

Layer-level knowledge storage. As shown in Figure 4 (GPT2-large and Llama-7B), SDS-selected neurons exhibit a strong concentration in deeper layers, particularly near the final transformer blocks (e.g., layers 26–29 in GPT2-large and 30–31 in Llama-7B). In contrast, baseline methods such as log-probability-based selection and its random variant distribute neurons more broadly across middle layers, which are typically associated with intermediate representations rather than final prediction formation. This suggests that static or weakly perturbed methods tend to capture correlational signals, whereas SDS more effectively isolates causally relevant neurons through dynamic verification.

Statistical analysis across six knowledge categories (Table 2 and Figure 5) reveals a clear functional division within Llama-7B's knowledge organization (similar results for GPT2-large are provided in Appendix B). FFN neurons are predominantly concentrated in mid-to-late and final layers across all tasks, suggesting they serve as stable storage components of factual knowledge. In contrast, attention neurons are more dispersed and task-dependent, indicating their role in dynamic routing and information selection rather than persistent storage. Overall, these findings support the hypothesis that SDS aligns with the model's intrinsic knowledge architecture, capturing causally relevant components rather than superficial performance-correlated activations.

Neuron-level knowledge analysis. To evaluate the extent of knowledge encoded in neurons identified by SDS, we perform a budgeted ablation study by

top10 important layers										
lang	f_{31}	f_{30}	f_{21}	f_{22}	f_{24}	f_{18}	f_{28}	f_{23}	f_{16}	f_{29}
capi	f_{31}	f_{30}	f_{17}	f_{18}	f_{19}	f_{21}	f_{29}	f_{28}	f_{23}	f_{22}
cnty	f_{31}	f_{30}	f_{18}	f_{28}	f_{22}	f_{23}	f_{19}	f_{20}	f_{17}	f_{21}
col	f_{31}	f_{30}	f_{29}	f_{20}	f_{22}	f_{21}	f_{27}	f_{19}	f_{25}	f_{17}
num	f_{31}	f_{24}	f_{26}	f_{29}	f_{17}	f_{30}	f_{28}	f_{22}	f_{20}	f_{27}
mon	f_{31}	f_{30}	f_{19}	f_{22}	f_{18}	f_{20}	f_{29}	f_{17}	f_{27}	f_{21}
lang	a_{23}	a_{18}	a_{31}	a_{21}	a_{16}	a_{26}	a_{28}	a_{25}	a_{19}	a_{22}
capi	a_{25}	a_{21}	a_{18}	a_{23}	a_{31}	a_{19}	a_{29}	a_{16}	a_{20}	a_{28}
cnty	a_{25}	a_{23}	a_{26}	a_{29}	a_{18}	a_{31}	a_{21}	a_{16}	a_{27}	a_{19}
col	a_{29}	a_{20}	a_{28}	a_{16}	a_{19}	a_{25}	a_{18}	a_{26}	a_{17}	a_{31}
num	a_{26}	a_{19}	a_{28}	a_{18}	a_{21}	a_{29}	a_{27}	a_{17}	a_{30}	a_{31}
mon	a_{23}	a_{21}	a_{19}	a_{16}	a_{28}	a_{29}	a_{30}	a_{17}	a_{18}	a_{20}

Table 2: Top 10 important FFN and Attention layers in Llama-7B.

zeroing out approximately the top 1% of neurons in Llama-7B and GPT2-large. In Llama-7B, we jointly intervene on 200 FFN neurons and 100 attention neurons selected by SDS, and measure the resulting changes in model predictions. Similar results for GPT-2 are provided in Appendix C.

As shown in Table 3, across six categories of factual knowledge tasks, the model's predicted probability for the correct answer exhibits an almost complete collapse. The decline in prediction probability generally exceeds 98%, with tasks such as Country, Month, and Number experiencing drops of over 99.5%, indicating an extremely pronounced downward trend. Meanwhile, the MRR also shows a consistent and significant decrease, with overall reductions ranging from 94% to 98%. This indicates that, at the ranking level, the model almost entirely loses its ability to distinguish the correct knowledge tokens, resulting in severely degraded performance.

These results indicate that, despite the large number of parameters and neurons in Transformer models, the prediction of specific factual knowledge depends on a small, budgeted subset of neurons that exert a decisive causal influence on internal computation and information flow. When this causally selected subset is jointly intervened upon, even though it constitutes only a negligible fraction of the overall network, the model's ability to retrieve and express knowledge degrades severely, leading to substantial failures in prediction and generalization. Moreover, these findings validate SDS as an effective approximate solver for budgeted causal

	lang	capi	cnty	col	mon	num
Prob	0.47	0.57	0.38	0.18	0.27	0.16
↓ Prob%	98.8	98.9	99.6	98.9	99.5	99.9
MRR	0.97	0.98	0.91	0.78	0.89	0.29
↓ MRR%	94.3	95.4	97.9	97.1	97.8	98.5

Table 3: Impact of zeroing key neurons on multi-type knowledge prediction in Llama-7B

neuron selection, demonstrating its capability to reliably identify causally critical neurons across architectural components. Collectively, our results provide strong empirical support for the hypothesis that factual knowledge in large language models is encoded in a sparse, centralized, and structurally organized manner.

6 Conclusion

We propose SDS, a static–dynamic selection framework for budgeted causal neuron selection, formulated as an approximation strategy for efficient causal attribution under constrained intervention budgets. The method first performs high-recall static screening to construct a compact candidate set, and then applies dynamic causal ablation within this reduced space to identify functionally critical neurons. This design significantly reduces the search space of the underlying combinatorial optimization problem while preserving causal fidelity.We evaluate SDS on GPT-2 large and Llama-7B across multiple factual knowledge benchmarks. Results show that SDS consistently outperforms existing attribution methods in terms of prob, logp, and MRR, and maintains stable performance across different model scales and intervention budgets. These improvements demonstrate that SDS effectively approximates causally important neuron subsets under limited computational resources.

Beyond performance gains, structural analysis reveals that neurons identified by SDS exhibit both cross-task consistency and task-specific variation across FFN and attention layers. This suggests that SDS not only improves attribution accuracy, but also aligns with the underlying functional organization of Transformer representations.Overall, SDS provides an efficient and reliable solution to budgeted causal neuron attribution, bridging combinatorial optimization and mechanistic interpretability through a static–dynamic decomposition strategy.

Limitations

Although SDS demonstrates strong performance in budgeted causal neuron attribution, its effectiveness is still partially constrained by the quality of the static candidate generation stage. Since the framework relies on a high-recall static filter to construct the candidate set, neurons with low static scores may be excluded even if they are causally important. While increasing the expansion ratio α can mitigate this issue, it introduces additional computational overhead, indicating a trade-off between recall and efficiency inherent to the budgeted optimization design.

In addition, SDS is currently based on single-neuron ablation, which assumes independence among neurons and does not explicitly model higher-order interactions or synergistic effects. However, knowledge in large language models may be distributed across groups of neurons or subnetworks, suggesting that single-point intervention may provide only an approximation of the true causal structure. Extending SDS to subnetwork- or path-level causal attribution remains an important but computationally challenging direction.

Furthermore, our evaluation is primarily conducted on factual knowledge tasks using autoregressive language models such as GPT-2 large and Llama-7B. The generalization of SDS to instruction-tuned models, multimodal architectures, and more complex reasoning tasks remains to be explored. Finally, due to its ability to identify and manipulate causally relevant neurons, SDS may introduce potential misuse risks in model editing or targeted intervention scenarios, highlighting the need for responsible use in mechanistic interpretability research.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Lihu Chen, Adam Dejl, and Francesca Toni. 2025. Identifying query-relevant neurons in large language models for long-form texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23595–23604.

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

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.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170.

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

Harshwardhan Fartale, Ashish Kattamuri, Rahul Raja, Arpita Vats, Ishita Prasad, and Akshata Kishore Moharir. 2025. Disentangling recall and reasoning in transformer models through layer-wise attention and activation analysis. *arXiv preprint arXiv:2510.03366*.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 30–45.

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.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, and Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36:17643–17668.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32.

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.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

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.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Leland Gerson Neuberg. 2003. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, and Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

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

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4932–4942.

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.

Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, and Joseph Bloom, et al. 2025. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, and Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4396–4406.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019b. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5797–5808.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *International conference on machine learning*, pages 10524–10533. PMLR.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240.

Zeping Yu and Sophia Ananiadou. 2024. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280.

A Neuron Ablation Analysis on GPT2-large across Attribution Methods

The behavior of different attribution methods under varying intervention budgets n in GPT2-large is consistent with that observed in Llama-7B, indicating strong cross-model robustness under the budgeted causal neuron selection setting. As shown in Figure 6, we compare SDS with log-probability-based methods and their random variants, and find that increasing the number of ablated neurons n leads to a consistent degradation in prediction probability, log probability, and MRR across all methods.

Notably, SDS induces a more pronounced degradation than all baselines across all budget levels, further confirming its ability to identify causally critical neurons in a more accurate approximation of the underlying optimization objective.

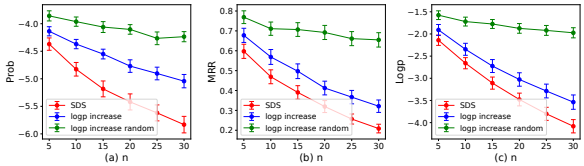

Figure 6: Neuron Ablation Analysis on GPT2-large across Attribution Methods.

B Top 10 FFN and Attention layers in GPT2-large and their corresponding heatmaps

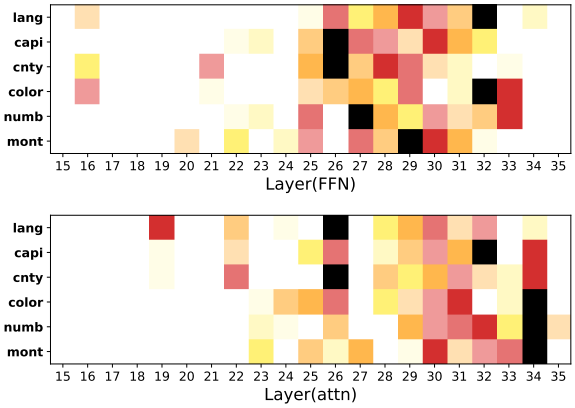

Figure 7: Heatmap of importance scores for the top-10 FFN and Attention layers in GPT2-large.

Table 4 presents a structural analysis of the top-10 key layers identified by SDS across different semantic knowledge tasks, including language, capitals, countries, colors, numbers, and months. The results show clear layer-wise specialization, where different tasks activate distinct subsets of FFN and attention layers, reflecting task-dependent functional allocation within the model.

	top10 important layers									
lang	f_{32}	f_{29}	f_{26}	f_{30}	f_{28}	f_{31}	f_{16}	f_{27}	f_{34}	f_{25}
capi	f_{26}	f_{30}	f_{27}	f_{28}	f_{31}	f_{25}	f_{29}	f_{32}	f_{23}	f_{22}
cnty	f_{26}	f_{28}	f_{29}	f_{21}	f_{25}	f_{27}	f_{30}	f_{16}	f_{31}	f_{33}
col	f_{32}	f_{33}	f_{29}	f_{16}	f_{27}	f_{26}	f_{25}	f_{28}	f_{31}	f_{21}
num	f_{27}	f_{33}	f_{25}	f_{30}	f_{28}	f_{32}	f_{31}	f_{29}	f_{23}	f_{22}
mon	f_{29}	f_{30}	f_{27}	f_{25}	f_{31}	f_{28}	f_{20}	f_{22}	f_{24}	f_{32}
lang	a_{26}	a_{19}	a_{30}	a_{32}	a_{29}	a_{22}	a_{31}	a_{28}	a_{34}	a_{24}
capi	a_{32}	a_{34}	a_{26}	a_{30}	a_{31}	a_{29}	a_{22}	a_{25}	a_{28}	a_{19}
cnty	a_{26}	a_{34}	a_{22}	a_{31}	a_{30}	a_{28}	a_{32}	a_{29}	a_{33}	a_{19}
col	a_{34}	a_{31}	a_{26}	a_{30}	a_{25}	a_{24}	a_{29}	a_{28}	a_{33}	a_{23}
num	a_{34}	a_{32}	a_{31}	a_{30}	a_{29}	a_{26}	a_{35}	a_{33}	a_{23}	a_{24}
mon	a_{34}	a_{30}	a_{33}	a_{32}	a_{27}	a_{25}	a_{31}	a_{23}	a_{26}	a_{29}

Table 4: Top 10 important FFN and Attention layers in GPT2-large.

	lang	capi	cnty	col	mon	num
Prob	0.03	0.04	0.05	0.02	0.02	0.01
↓ Prob%	96.16	88.79	90.56	90.87	87.57	73.5
MRR	0.23	0.32	0.33	0.18	0.21	0.17
↓ MRR%	93.0	81.0	83.8	87.8	84.9	67.1

Table 5: Impact of zeroing key neurons on multi-type knowledge prediction in GPT2-large.

Figure 7 further visualizes the distribution of these key layers across the depth of GPT2-large. We observe that SDS-identified neurons are predominantly concentrated in the middle-to-late layers of the model, indicating that deeper layers play a central role in encoding and retrieving semantic knowledge under the budgeted causal selection framework.

C Impact of Critical Neuron Intervention on Knowledge Prediction in GPT2-large

Table 5 presents the effects of zero-out interventions on the top 1% of neurons identified by SDS in GPT2-large across multiple factual knowledge tasks. We observe that ablating this small budgeted causal subset leads to severe performance degradation across all tasks, consistent with the results observed in Llama-7B. Across six tasks, predic-

tion probability and ranking performance (MRR) exhibit substantial drops following intervention. Language-related tasks show the largest degradation, while even more stable tasks such as Number still experience significant performance loss. Overall, probability decreases range from approximately 70% to over 96%, and MRR declines consistently across tasks, indicating a near-complete breakdown in knowledge retrieval capability.

These results demonstrate that the neuron subset identified by SDS plays a critical causal role in model prediction behavior. Despite constituting only a small fraction of the overall network, intervening on this budgeted causal subset leads to system-wide performance collapse, supporting the hypothesis that factual knowledge in large language models is encoded in a sparse and highly structured manner.