Do All Autoregressive Transformers Remember Facts the Same Way? A Cross-Architecture Analysis of Recall Mechanisms

Anonymous ACL submission

Abstract

Understanding how Transformer-based language models store and retrieve factual associations is critical for improving interpretability and enabling targeted model editing. Prior work, primarily on GPT-style models, has identified MLP modules in early layers as key contributors to factual recall. However, it remains unclear whether these findings generalize across different autoregressive architectures. To address this, we conduct a comprehensive evaluation of factual recall across several models—including GPT, LLaMA, Qwen, and DeepSeek—analyzing where and how factual information is encoded and accessed. Consequently, we find that Qwen-based models behave differently from previous patterns: attention modules in the earliest layers contribute more to factual recall than MLP modules. Our findings suggest that even within the autoregressive Transformer family, architectural variations can lead to fundamentally different mechanisms of factual recall.

1 Introduction

Transformer-based language models are trained in an autoregressive manner, generating the next token sequentially based on previous tokens. These models exhibit strong language understanding due to their scale and self-attention mechanism (Vaswani et al., 2017). Beyond simple sentence generation, they internalize factual associations, typically expressed in knowledge triples of the form $t = (s, r, o)$, where s is the subject, r is the relation, and o is the object, and effectively recall them during inference (Petroni et al., 2019).

Factual association recall refers to a model's ability to generate the correct object given a subject–relation prompt by leveraging internally stored factual knowledge. One line of research on factual recall investigates whether specific properties are embedded within the internal representations of language models, by tracing information flow through causal analysis (Meng et al., 2022a,b; Geva et al., 2023). These approaches are widely used in applications such as knowledge editing to identify where factual associations are primarily stored and to enable targeted modifications, rather than editing the entire network.

However, these prior findings have focused only on GPT-family architectures. This observation raises the question of whether *such findings generalize across different autoregressive architectures*. To address this gap, we conduct a comprehensive, layer-wise and module-wise (Attention vs. MLP) quantitative evaluation of factual association recall—examining where and how factual information is stored and retrieved across a range of autoregressive Transformer models, including GPT (Radford et al., 2019; Brown et al., 2020; Wang and Komatsuzaki, 2021), LLaMA (Grattafiori et al., 2024), Qwen (Yang et al., 2024), and DeepSeek (Guo et al., 2025) [1]. Our evaluation and analysis reveals *notable architecture-specific differences*—particularly in the Qwen-based architectures—in where and how factual associations are recalled.

In our evaluation, we extend the experimental frameworks of Meng et al. (2022a) and Geva et al. (2023), applying causal tracing across a diverse set of autoregressive Transformer models to perform the following experiments, as showed in Figure 1.

1. Restoration effects. We measure the change in inference probability between clean and corrupted inputs at the last subject token position, following causal tracing protocols from prior work. This allows us to analyze the sensitivity of model predictions across all layers and modules (Attention vs. MLP), providing insight into where factual

[1] Evaluation results are based on 17 models across target architectures; For brevity, we present results for four representative models in the main text, with the remaining results included in Appendix A.

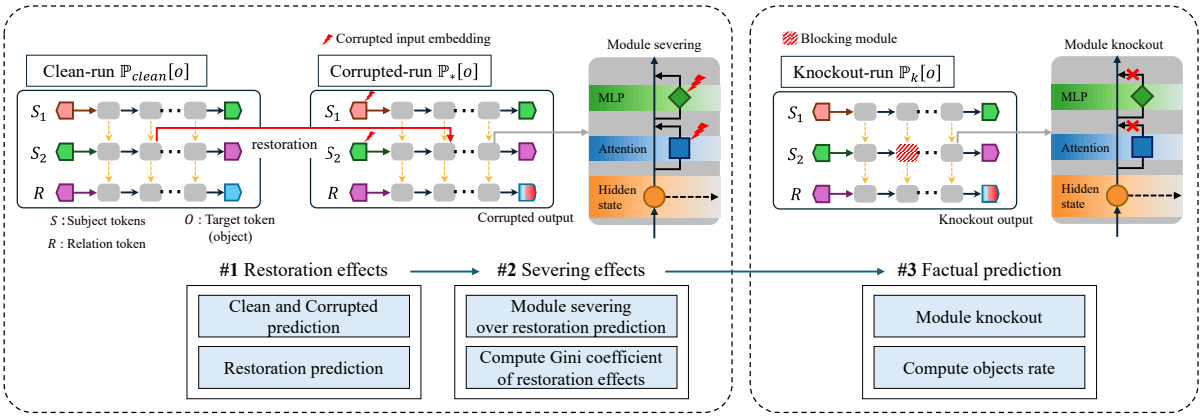

Figure 1: **Overview of Our Evaluation.** The pipeline consists of three main stages: (1) restoration effects, (2) severing effects, and (3) factual prediction. Each stage applies a distinct intervention strategy to evaluate the contribution of the Attention and MLP modules to factual association recall.

associations are most affected by input corruption.

2. Severing effects. We perform selective severing either the Attention or MLP module at each layer during inference. We then evaluate how much each module contributes to the model's ability to generate the correct object in factual association prompts. In addition, we compute the Gini coefficient over inference contributions to assess how structurally localized the causal effects are within each module.

3. Factual prediction. We refine the object prediction evaluation method used in Geva et al. (2023), which relies on string matching between predicted tokens and BM25-selected candidates. Noting its insensitivity to semantically equivalent outputs, we introduce a semantic similarity–based evaluation metric, which considers a predicted token correct if its embedding similarity to a candidate token exceeds a fixed threshold. This allows for more robust evaluation of model outputs, especially in architectures that generate more varied but semantically equivalent expressions.

Results. This study highlights the following:

1. Our evaluation confirms that, in GPT-based models, MLP modules in the early layers play a key role in storing factual associations. This finding supports prior work (Meng et al., 2022a; Geva et al., 2023), demonstrating that their conclusions are applicable to this class of architectures.

2. However, we identify key factors indicating that Qwen-based models show larger changes in inference probability within the *early Attention layers*, unlike GPT-based models. In Qwen-based models, factual associations are more concentrated in the Attention module than in the MLP. This finding is further supported by Gini coefficient analysis, which shows that the Attention modules are where most of tqhe important effects are focused—highlighting their key role in factual recall in this model.

3. Our semantic similarity–based evaluation further confirms the difference between GPT-based and Qwen-based models. In Qwen-based models, Attention modules consistently contribute more to factual inference than MLPs, even when the model's outputs vary in wording.

These findings have practical implications for deploying Transformer models in real-world scenarios, *where knowing which layers and modules store factual knowledge is important* for effective targeted model editing, reliable interpretability, and knowledge-intensive tasks.

2 Notation of Autoregressive Transformer

We begin by outlining the fundamental architecture of autoregressive Transformer models (Vaswani et al., 2017), before analyzing the recall mechanisms of factual associations in representative models such as GPT, LLaMA, Qwen, and DeepSeek. These models are designed such that each token can perform the Attention mechanism only over tokens that precede it.

An autoregressive Transformer model $M : X \to Y$ operates over a vocabulary V, and given an input token sequence $x = [x_1, ..., x_T] \in X$, produces a probability distribution $y \in Y \subset \mathbb{R}^{|V|}$ for the next token. This distribution reflects the model's estimation of the next token given the preceding context.

Internally, each token x_i is mapped to a hidden state vector $h_i^{(l)}$ at layer l. The initial hidden state $h_i^{(emb)}$ is computed as the sum of the token em-

bedding and positional embedding, as defined in Equation 1.

$$h_i^{(emb)} = emb(x_i) + pos(i) \in \mathbb{R}^{d_{model}} \quad (1)$$

The hidden state is iteratively updated through layers using Attention and MLP modules. At layer l, the hidden state of the i-th token is :

$$h_i^{(l)} = h_i^{(l-1)} + a_i^{(l)} + m_i^{(l)} \quad (2)$$

The Attention output $a_i^{(l)}$ incorporates representations from previous tokens via self-attention :

$$a_i^{(l)} = attn^{(l)}(h_1^{(l-1)}, ..., h_i^{(l-1)}) \quad (3)$$

The MLP output applies a non-linear transformation to the residual input :

$$m_i^{(l)} = W_{proj}^{(l)} \sigma(W_{fc}^{(l)} \gamma(a_i^{(l)} + h_i^{(l-1)})) \quad (4)$$

where γ denotes a normalization, and σ is a non-linear activation.

3 Structural Insights into Factual Recall

We conduct a quantitative analysis of how factual associations are stored across various autoregressive models through a unified evaluation. The results show that Qwen and DeepSeek exhibit patterns that differ from previously reported trends. Specifically, the early Attention layers corresponding to the position of the last subject token make significant contributions to the storage of factual associations. These findings suggest that architectural differences may influence the mechanisms underlying factual association storage.

We frame our empirical study around the following research questions, and describe our approach.
#1. In various autoregressive Transformer models, which modules, layers, and token-position-specific activations contribute to the recall of specific factual associations, and to what extent do they influence the model's output?

- Following the causal tracing methodology introduced in ROME (Meng et al., 2022a), we conduct **restoration effects** and **severing effects** experiments across a range of autoregressive Transformer models. Through these experiments, we quantitatively measure the Average Indirect Effect (AIE) of internal activations.

#2. Experiments conducted across various models reveal an inconsistency between the results of the restoration effects experiment and the severing effects experiment in the Qwen and DeepSeek models, raising the question: what factors might have caused this discrepancy?

- To investigate the underlying cause of this discrepancy, we compute the **Gini coefficient** over the distribution of AIE values to quantify the degree of concentration across layers, thereby identifying which layers contribute most to factual association recall.

- Subsequently, we apply the severing effects experiment to the layers within the Attention modules of the Qwen and DeepSeek models that exhibit the highest AIE concentration. However, severing these layers does not result in a significant reduction in AIE. The inconsistency between the two experimental results is consistently observed.

#3. These findings suggest that the severing effects may have failed to sufficiently suppress the activations of the Attention module. Does this not indicate the need for a more precise intervention method?

- Experimental results reveal that the severing effects method (as proposed in Meng et al. (2022a)) has a critical limitation in accurately reflecting the contribution to factual association recall, as it fails to sufficiently suppress the influence of the attention module.

- To address this limitation, we adopt the methodology proposed by Geva et al. (2023) and apply a **knockout** technique that directly blocks the output of each module. In addition, we evaluate factual prediction performance using the objects rate instead of AIE, incorporating semantic similarity by computing the cosine similarity between Sentence-BERT (Reimers and Gurevych, 2019) embeddings, rather than relying on exact string matching. Through this approach, we empirically confirm that the early layers of the Attention module in the Qwen and DeepSeek models contribute to the factual recall.

We present a novel finding that, in Qwen-based models, the early layers of the Attention module contribute more to factual association recall at the

Model	#Layers	#Parameters
GPT-2-XL	48	1.5B
LLaMA-3.2-1B	16	1B
Qwen-2.5-1.5B	28	1.5B
DeepSeek-R1 Distill-Qwen-1.5B	28	1.5B

Table 1: Architectural details of autoregressive Transformer models (B = Billion parameters).

last subject token position than the MLP module. This contrasts with prior work (Meng et al., 2022a; Geva et al., 2023) that identifies the MLP as the primary site for factual association storage. Our result suggests that the localization of factual associations can shift depending on model architecture, highlighting the need for broader comparative studies across Transformer families.

4 Evaluation Methods

We build on and modify the implementations provided by Meng et al. (2022a) and Geva et al. (2023) to suit our evaluation framework.

Model Selection. We evaluate a variety of autoregressive Transformer models, as summarized in Table 1. Additional experimental targets and evaluation results are provided in the Appendix A.

Dataset & Prompting. We employ the COUNTERFACT dataset (Meng et al., 2022a), which is specifically designed to test factual associations stored in language models. This dataset consists of a diverse collection of prompts that encode factual associations.

To examine how each model processes the object in a factual association, we input only the subject and relation from a knowledge tuple $t = (subject, relation, object)$, excluding the object. This setup allows us to evaluate the model's ability to accurately predict the object based on the given context. For performance evaluation, we sample 100 factual sentences in which the object was successfully predicted, and apply the evaluation methodology described in the next subsection. Throughout all experiments, model layers are indexed from 0 to $L-1$, and our analysis focuses on the last subject token position, which is critical for triggering factual recall within the model.

4.1 Restoration Effects

Our first experiment **#1** aims to quantitatively analyze how specific modules, layers, and token-position-specific activations in Transformer-based language models contribute to the storage and recall of factual associations.

The analysis assumes two baseline execution settings. The **clean-run** executes the model on an uncorrupted input, yielding accurate factual predictions. In contrast, the **corrupted-run** injects noise into the subject representation to degrade recall performance. Specifically, we perturb the subject embedding by adding noise with a magnitude of $v = 3\sigma_{sub}$, where σ_{sub} is the standard deviation of subject embeddings collected from the dataset.

Based on this setup, we quantify the restoration effects by restoring the activation of a module at a specific layer to its clean state within the corrupted run. This setup enables us to measure the causal contribution of the target component to factual association recall.

Under this setup, we use the metric Indirect Effect (IE) to quantify the degree to which a given component contributes to factual recall. IE is computed as the difference in prediction probability of the correct token o between the restored and non-restored cases (Meng et al., 2022a):

$$IE = \mathbb{P}_{*,\text{clean } h_i^{(l)}}[o] - \mathbb{P}_*[o] \tag{5}$$

where $\mathbb{P}_*[o]$ denotes the prediction probability of o in the corrupted run, and $\mathbb{P}_{*,\text{clean } h_i^{(l)}}[o]$ is the corresponding probability when the hidden state $h_i^{(l)}$ is restored to its clean value. In this experiment, IE is used to quantify the causal contribution of the target component, where a higher IE indicates that the activation of the module at the given layer has a greater impact on factual recall.

Finally, the Average Indirect Effect (AIE) aggregates IE across multiple prompts, capturing the degree to which a given component consistently supports factual recall across diverse contexts.

4.2 Severing Effects

Our second experiment **#2** applies the severing effects to quantitatively analyze the extent to which AIE-based contributions are concentrated in either the Attention or MLP modules. Specifically, similar to the restoration effects, we start from a hidden state in which activations have been restored, but selectively replace only the activation of the target module (e.g., Attention or MLP) with its corrupted counterpart, thereby severing the information flow through that module. If the AIE value significantly

decreases when the activation of a specific module at a certain layer is corrupted, it suggests that the corresponding activation plays a critical role in factual association recall.

The restoration and severing effects experiments yield diverging results, making it difficult to consistently interpret the contribution of each module. To address this discrepancy, we compute the Gini coefficient (Dorfman, 1979) over the AIE distribution to quantify the concentration of contributions across layers. The Gini coefficient is defined as follows:

$$G_A = \frac{\sum_i \sum_j |AIE'_i - AIE'_j|}{2L \sum_i AIE'_i} \tag{6}$$

where L is the number of total layers, and AIE' denotes the normalized, layer-level aggregated AIE values.

We compute the Gini coefficient to identify the layer with concentrated AIE, and then apply the severing effects to evaluate whether the AIE reduction aligns proportionally with the restoration effects.

4.3 Factual Prediction

Our third experiment **#3** aims to quantitatively evaluate how the Attention and MLP modules inject factual association into the subject representation and functionally contribute to factual prediction. To this end, we execute a **knockout-run**, which sequentially applies knockout interventions to each layer $l = 0, \ldots, L-1$. Specifically, we zero out the updates at the last subject token position in both the Attention and MLP modules across five consecutive layers. The intervention is applied over the range $\tilde{l} = l, \ldots, \min\{l+4, L-1\}$, and the model's top-$k$ output tokens under these intervention conditions are collected. These tokens are subsequently used as the set T, which serves as the basis for factual prediction evaluation.

Prior to evaluation, we construct a candidate object set O corresponding to each subject. We retrieve relevant paragraphs from Wikipedia using BM25 (Robertson et al., 1995; Geva et al., 2023), followed by tokenization and the removal of stopwords and subword fragments. The resulting candidate set consists of non-common tokens that frequently co-occur with the subject and can be considered plausible object expressions.

Language models with a relatively large number of parameters often produce outputs that are semantically accurate but lexically differ from the canonical object tokens. As a result, conventional string match-based metrics significantly underestimate the objects rate. To mitigate this limitation, we adopt a semantic similarity-based evaluation method using Sentence-BERT (Reimers and Gurevych, 2019).

Therefore, instead of relying on exact string matching, we compute the cosine similarity between the Sentence-BERT embeddings of the generated tokens and the object candidates. Tokens that exceed a predefined similarity threshold τ with any element in O are regarded as semantically valid. This design allows for robust and generalizable evaluation of factual prediction by accounting for lexical variation, improving upon string match-based methods used in prior work (Geva et al., 2023).

The objects rate is defined as the average proportion of semantically valid tokens among the top-k outputs for each subject:

$$ObjectsRate = \frac{|O_\tau|}{|T|} \times 100 \tag{7}$$

where T denotes the set of top-k tokens (with $k = 50$) generated by the model for a given subject, and O_τ is defined as:

$$O_\tau = \{t \in T \mid \exists o \in O,\ \mathrm{sim}(t, o) \geq \tau\} \tag{8}$$

Here, O is the candidate object set for each subject, and $\mathrm{sim}(t, o)$ denotes the cosine similarity between the generated token t and the object candidate o. The similarity threshold is set to $\tau = 0.7$[2], and the objects rate is averaged across all subjects to yield a generalized measure of factual prediction performance.

5 Evaluation Results

5.1 Restoration Effects Analysis

To analyze the components contributing to the recall of factual associations, we use the restoration effects experiment to compute the Average Indirect Effect (AIE). The experiments are conducted on autoregressive Transformer-based models including GPT, LLaMA, Qwen, and DeepSeek, with a particular focus on the last subject token, which plays a crucial role in factual association recall.

Figure 2 shows the results of the restoration

[2]Detailed in Appendix B.

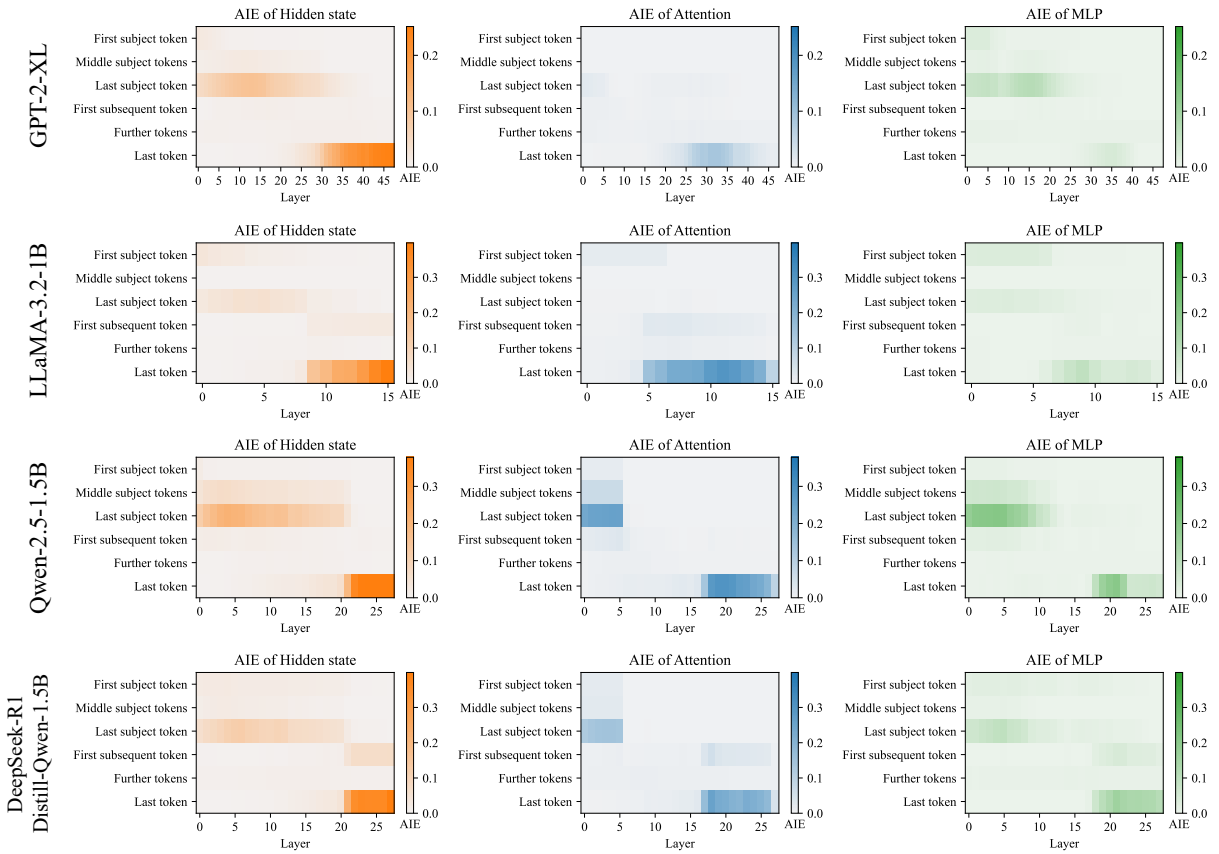

Figure 2: Restoration effects across multiple autoregressive Transformer models.

effects experiment, visualizing the layer-wise and module-wise AIE distribution across all token positions for each model. In GPT and LLaMA, AIE values are elevated in the early MLP layers, reflecting a canonical pattern of factual recall (Meng et al., 2022a). By contrast, Qwen and DeepSeek exhibit notably high AIE values in the early layers of the Attention module, suggesting that, unlike GPT and LLaMA, they store factual associations in the early Attention layers.

5.2 Severing Effects Analysis

To quantitatively analyze whether AIE-based contributions are more concentrated in the Attention or MLP modules, we perform a severing effects experiment in which we remove the outputs of specific layers and measure the resulting changes in AIE. As prior results (Figure 2) indicate that the early layers of both modules exhibit relatively high AIE values, we design the experiment to target layers 0 through 15 for each module.

Figure 3 shows the results of the severing effects experiment, focusing on layers with high AIE values identified in the previous restoration effects analysis. In GPT and LLaMA, the AIE values are elevated in the early MLP layers, and severing these layers leads to a substantial reduction in AIE. By contrast, in Qwen and DeepSeek, although the early Attention layers exhibit high AIE values, severing these layers results in only a minimal decrease. This indicates that while the Attention module appears to contribute significantly in the restoration effects experiment, its influence is not effectively suppressed in the severing effects setting.

To quantitatively explain this discrepancy, we compute the Gini coefficient of the layer-wise AIE distribution, as shown in Figure 4. The results show that in the Qwen and DeepSeek models, the AIE contribution of the Attention module is highly concentrated in the fourth layer. In contrast, in the GPT and LLaMA models, contributions from both the Attention and MLP modules are more evenly distributed. Based on this analysis, we conduct an additional experiment that selectively severs the layers with the highest concentration of contributions in each module. The results are summarized in Table 2. Severing the MLP layers leads to a clear drop in AIE, while severing the Attention layers yields only a minimal reduction. This outcome can

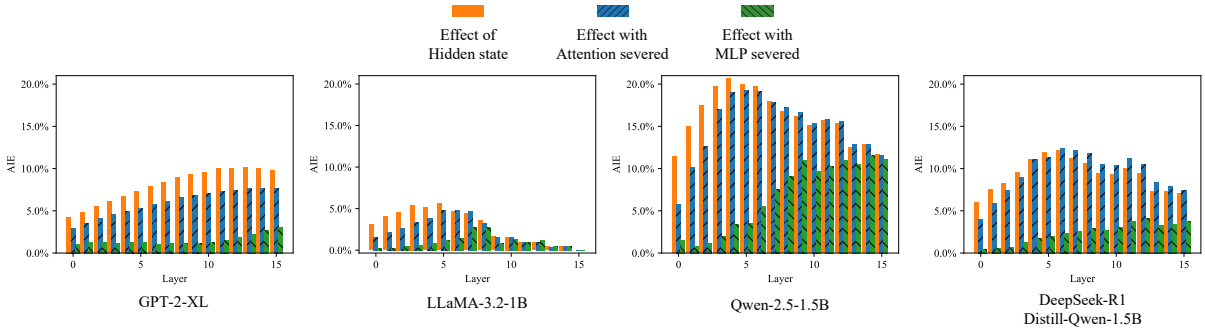

Figure 3: Severing effects across multiple autoregressive Transformer models (restricted to layers 0–15 to verify whether severing early layers significantly reduces AIE, consistent with the concentration observed in the restoration effects experiment).

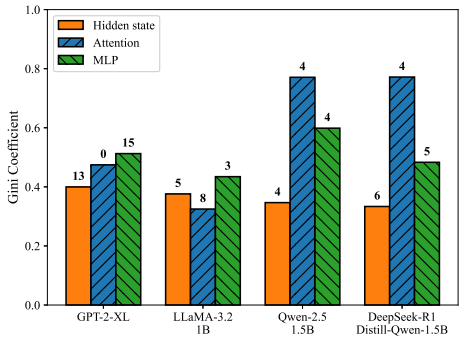

Figure 4: Gini coefficient-based concentration analysis in Attention and MLP modules across autoregressive Transformer models (numbers above bars indicate the layer with the highest concentration).

Model	Attention (Drop rate)	MLP (Drop rate)
GPT-2-XL	31.32%	69.52%
LLaMA-3.2-1B	10.82%	89.40%
Qwen-2.5-1.5B	8.25%	83.55%
DeepSeek-R1 Distill-Qwen-1.5B	-0.14%	83.60%

Table 2: Effect of severing highly concentrated layers selected by the Gini coefficient on AIE. A larger drop rate indicates greater layer influence (for Qwen and DeepSeek, severing Attention layers with high Gini scores results in only a small drop).

be attributed to the structural characteristics of the Attention mechanism. In the Attention module, information can propagate through alternative paths even when specific routes are blocked, which mitigates the impact of severing a particular layer (Elhage et al., 2021). In contrast, the MLP module operates independently on each token, and severing it directly disrupts the flow of information. These structural differences account for the observed disparity in AIE reduction between the two modules in the severing effects experiment (Geva et al., 2021).

5.3 Factual Prediction Analysis

To more precisely evaluate the contribution of the Attention module, we conduct a factual prediction experiment using a knockout approach that completely blocks the output of the target module.

Figure 5 shows the change in objects rate when a knockout is applied to each layer of the Attention and MLP modules. In GPT and LLaMA, blocking the early MLP layers leads to a substantial drop in objects rate, whereas in Qwen and DeepSeek, blocking the early Attention layers also results in a substantial drop. These results suggest that GPT and LLaMA primarily store factual associations in the early layers of the MLP module, while Qwen and Qwen-based DeepSeek store a substantial portion of such associations in the early layers of the Attention module, which plays a critical role in the recall of factual associations. These findings indicate that the factual prediction experiment provides a more appropriate and reliable means of assessing the contribution of the Attention module than the severing effects experiment.

6 Practical Implications

Our study highlights important differences in how language models store and recall facts, depending on their architecture, which have important practical implications. For example, in Qwen-based architectures, where factual recall is concentrated in early Attention layers, knowledge editing methods (Meng et al., 2022a,b; Li et al., 2024; Fang et al., 2024) should adapt accordingly: targeting Attention modules rather than MLPs. Similarly, interpretability tools (Clark et al., 2019; Belinkov, 2022) and attribution analyses (Vig, 2019)

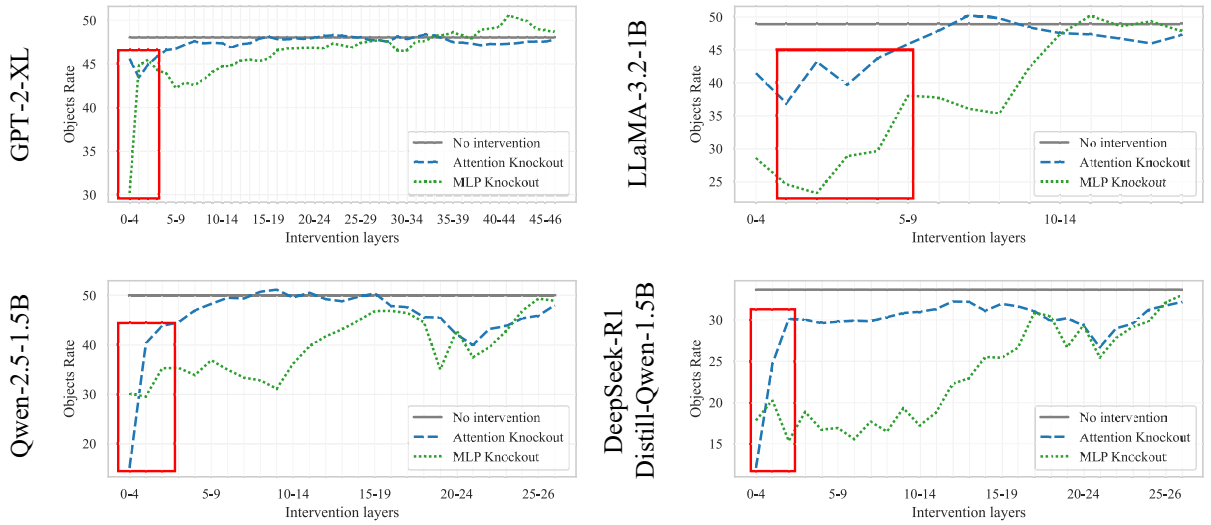

Figure 5: Factual prediction evaluation after knockout of Attention and MLP outputs across autoregressive Transformer models. Red boxes indicate points where blocking either the Attention or MLP module in early layers causes a substantial drop in objects rate, suggesting that the corresponding module plays a critical role in factual association recall.

should focus on these components to better trace factual reasoning. These insights can also guide architecture-aware model compression (Sanh et al., 2020), where preserving key Attention layers may help retain factual knowledge.

7 Related Work

Language models go beyond predicting sequences based solely on frequently occurring word patterns, as they internally store and utilize factual associations (Petroni et al., 2019; Jiang et al., 2020; Roberts et al., 2020). Early research on factual association recall focuses on probing methods, which train classifiers on frozen representations to evaluate whether specific properties are embedded within the internal representations of language models (Ettinger et al., 2016; Adi et al., 2016; Belinkov et al., 2017; Hupkes et al., 2018; Conneau et al., 2018; Elazar et al., 2021). However, such approaches are limited in explaining how these properties functionally influence model predictions (Belinkov, 2022).

To address these limitations, recent work turns to causal analysis, which estimates the functional contributions of internal components by applying counterfactual interventions (Vig et al., 2020; Pearl, 2022). A prominent method that applies this concept to hidden representation analysis is causal tracing, which enables the layer-level and module-level attribution of factual predictions (Meng et al., 2022a). Recent work actively explores causal tracing, which now serves as a central analytical framework for model interpretability (Dai et al., 2022; Mohebbi et al., 2023; Hase et al., 2023; Geva et al., 2023; Dar et al., 2023) and knowledge editing (Meng et al., 2022a,b; Li et al., 2024; Fang et al., 2024).

8 Conclusion

We conducted a comprehensive evaluation of factual association recall to examine whether prior findings based on GPT models generalize across a broader range of autoregressive Transformer architectures. To this end, we designed our evaluation methods to highlight structural differences in factual recall behavior across models. Our results reveal that Qwen-based models behave differently from previously reported patterns: factual recall is more strongly concentrated in the Attention modules, rather than the MLPs. Our findings have important implications for the deployment of Transformer models in real-world applications. In particular, understanding which layers and modules are responsible for storing factual knowledge is essential for improving performance in knowledge-intensive tasks, enabling more precise model editing, and supporting interpretability.

Limitations

While we analyze where factual associations are stored across a range of autoregressive Transformer models, including GPT, LLaMA, Qwen, and DeepSeek, the internal mechanisms by which these associations are recalled during inference remain insufficiently understood. Furthermore, despite observing model-specific recall patterns, how factual associations are dynamically accessed across different architectures remains unclear. In addition, although our results indicate that the storage locations of factual associations may vary depending on the model architecture, the interpretation of such structural differences is limited, warranting further investigation. Addressing these limitations is crucial for applications such as knowledge editing across existing and emerging Transformer-based architectures.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*.

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.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

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

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, Toronto, Canada. Association for Computational Linguistics.

Robert Dorfman. 1979. A formula for the gini coefficient. *The review of economics and statistics*, pages 146–149.

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

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.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st workshop on evaluating vector-space representations for nlp*, pages 134–139.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2024. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355*.

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

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

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*.

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.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

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.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024. Pmet: Precise model editing in a transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18564–18572.

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

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. *arXiv preprint arXiv:2301.12971*.

Judea Pearl. 2022. Direct and indirect effects. *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 373.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, and 1 others. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Victor Sanh, Thomas Wolf, and Alexander Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in neural information processing systems*, 33:20378–20389.

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. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Ben Wang and Aran Komatsuzaki. 2021. Gpt-j-6b: A 6 billion parameter autoregressive language model.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

A Additional Experimental Analysis

We extend our experiments to autoregressive Transformer models with varying parameter sizes, as listed in Table 3, and conduct additional evaluations accordingly.

A.1 Resotration Effects Analysis

We extend the restoration effects experiment by conducting additional analyses on autoregressive Transformer models with varying parameter scales.

We show the restoration effects results separately for each model family. Figure 6 shows the results for the GPT family, and Figures 7, 8, and 9 show the results for the LLaMA, Qwen, and DeepSeek families, respectively.

First, the analysis of GPT-family models reveals that the MLP modules in the early layers at the last subject token play a primary role in storing factual associations. Notably, GPT-2 Large exhibits a distinct pattern, where the Attention modules in the early layers also make a significant contribution to recalling factual associations.

Second, the analysis of LLaMA-family models reveals that, similar to the GPT family, the MLP modules in the early layers at the last subject token play a major role in storing factual associations.

Third, the analysis of Qwen-family models reveals that both MLP and Attention modules contribute to the storage of factual associations primarily in the early layers at the last subject token. This suggests that Qwen models leverage multiple modules within the early layers jointly in the process of storing factual associations.

Finally, the analysis of DeepSeek-family models reveals that the pattern of factual association storage varies depending on the base model used for distillation. Qwen-based models store factual associations in the early layers at the last subject token through both Attention and MLP modules, whereas LLaMA-based models show no clear contribution from the Attention modules at the same position. Furthermore, the distillation models exhibit factual association recall structures similar to their base models and display generally consistent patterns.

These results suggest that the location of factual association storage may vary depending on structural differences in model architecture.

Prior work, such as ROME (Meng et al., 2022a), suggests that in models with a small number of layers, factual associations may be stored in the early layers of the attention module at the position of the last subject token. However, our experimental results show that clear activation is not observed in the attention module at the last subject token position in models with a small number of layers, such as GPT-Small, GPT-Medium, and LLaMA-3.2–3B. These findings suggest that the contribution of the attention module to factual association storage is not determined solely by the number of layers, but may also be influenced by other architectural factors.

Model	#Layers	#Parameters
GPT-2-Small	12	0.124B
GPT-2-Medium	24	0.335B
GPT-2-Large	36	0.774B
GPT-J-6B	28	6B
LLaMA-3.1-8B	32	8B
LLaMA-3.2-3B	28	3B
Qwen-2.5-0.5B	24	0.5B
Qwen-2.5-3B	36	3B
Qwen-2.5-7B	28	7B
Qwen-2.5-14B	48	14B
DeepSeek-R1 Distill-Qwen-7B	28	7B
DeepSeek-R1 Distill-LLaMA-8B	32	8B
DeepSeek-R1 Distill-Qwen-14B	48	14B

Table 3: Architectural configurations of autoregressive Transformer models with varying parameter scales (B = billion parameters).

A.2 Severing Effects Analysis

We extend the original severing effects experiment, which was limited to layers 0-15, to cover all layers, and perform additional analyses on autoregressive Transformer models with varying parameter scales.

Figure 10 shows the results of the extended severing effects experiments conducted across the full layer range. Figures 11, 12, 13, and 14 show the outcomes for the GPT, LLaMA, Qwen, and DeepSeek families, respectively. The analysis of all model families — including GPT, LLaMA, Qwen, and DeepSeek — reveals that severing the MLP modules in the early layers at the last subject token leads to a substantial drop in AIE, indicating their critical role in factual association recall. Notably, while Qwen and DeepSeek models exhibit high AIE values for the early Attention modules at the last subject token in the severing effects experi-

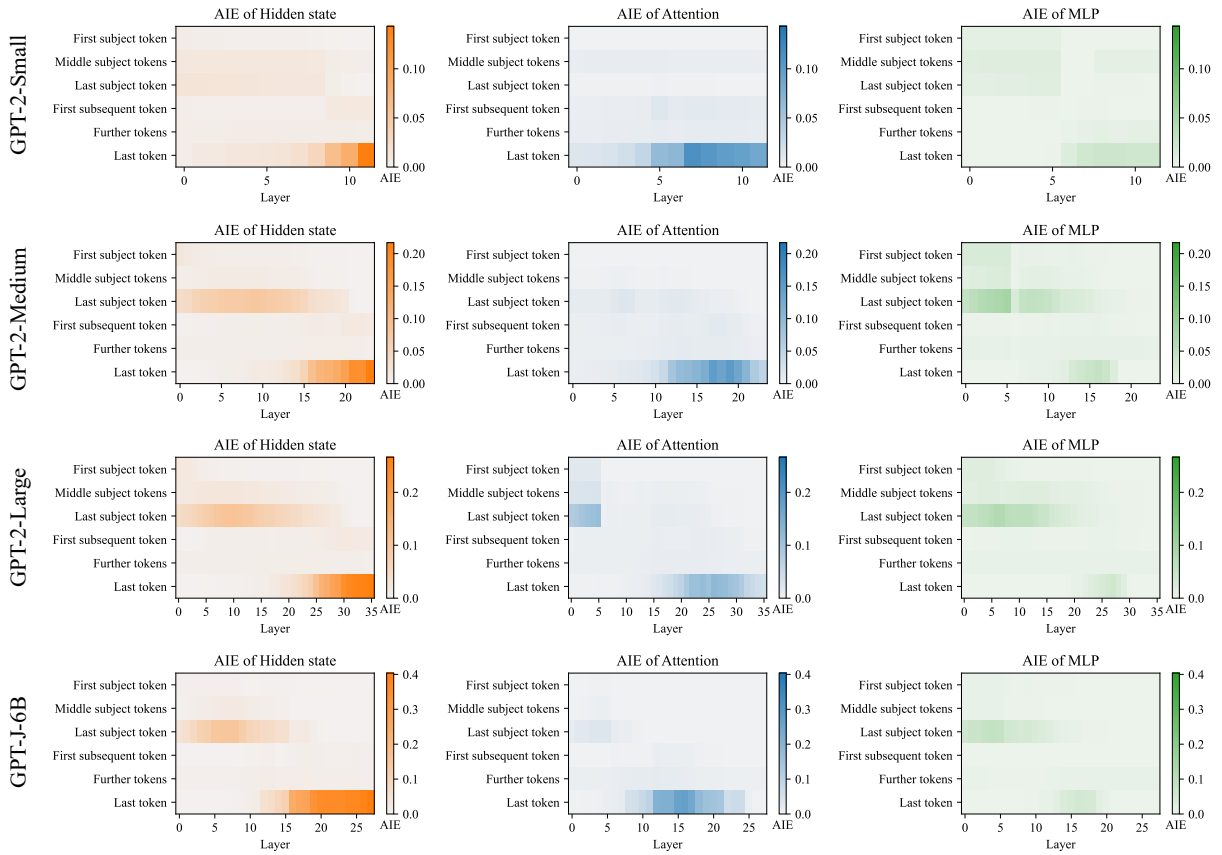

Figure 6: Restoration effects in GPT-family autoregressive Transformer models with varying parameter scales.

ments, the severing effects experiments show that severing these Attention modules does not result in as large a decrease in AIE as observed for the MLP modules.

Based on these results, we extend the previous Gini coefficient-based concentration experiments by conducting additional analyses across a range of autoregressive Transformer models to quantitatively examine the distribution characteristics of AIE within the Attention and MLP modules. Figure 15 shows the results of the Gini coefficient-based concentration analysis. Most GPT-family and LLaMA-family models, except for GPT-2-Small and GPT-2-Large, exhibit a relatively uniform distribution of AIE across the Attention and MLP modules. In contrast, Qwen-family and DeepSeek-family models show a concentrated distribution of AIE within specific layers of the Attention module, indicating that factual associations tend to be stored in a more concentrated manner within certain layers of the Attention module.

A.3 Factual Prediction Analysis

We extend the previous factual prediction experiments by conducting additional evaluations on a variety of autoregressive Transformer models.

The factual prediction results are presented separately for each model family. Figure 16 presents the results for the GPT family, and Figures 17, 18, and 19 correspond to the LLaMA, Qwen, and DeepSeek families, respectively.

Most GPT-family models exhibit a substantial drop in objects rate when the early layers of the MLP module are blocked, suggesting that these layers are key locations where factual associations are recalled.

Within the LLaMA family, LLaMA 3.1-8B shows a similar pattern to GPT models, with a significant objects rate drop when early MLP layers are blocked, indicating that factual associations are primarily recalled in that region. In contrast, LLaMA 3.2-3B shows a clear drop in objects rate when early Attention layers are blocked, implying that factual recall primarily occurs in the early Attention module.

Most models in the Qwen family also demonstrate that early Attention layers play a central role in factual association recall, as evidenced by marked decreases in objects rate when these layers are blocked.

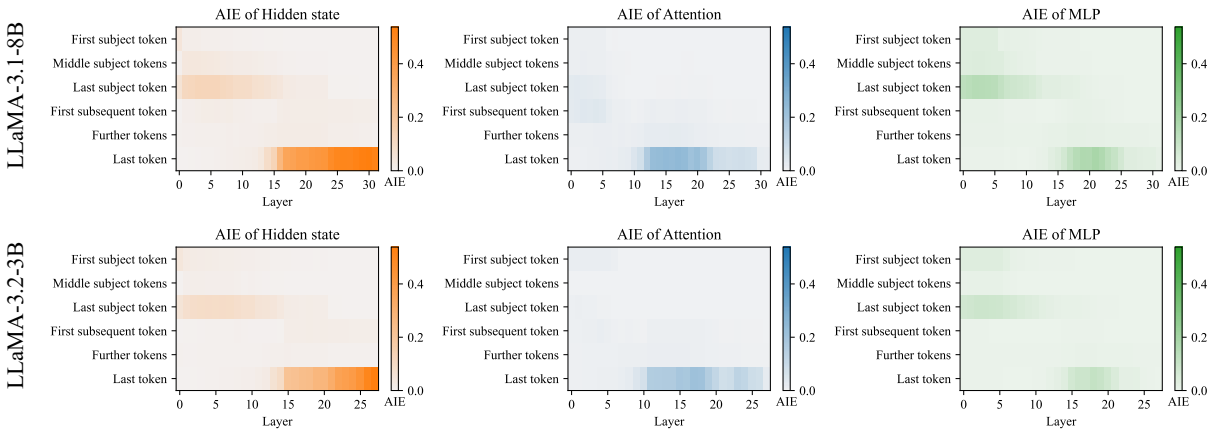

Figure 7: Restoration effects in LLaMA-family autoregressive Transformer models with varying parameter scales.

For the DeepSeek family, DeepSeek-R1-Distill-LLaMA-8B exhibits factual recall concentrated in the early MLP layers, while DeepSeek-R1-Distill-Qwen-14B shows that both early Attention and MLP layers contribute to factual recall. Meanwhile, DeepSeek-R1-Distill-Qwen-7B produces inconsistent results, indicating the need for further investigation in future work.

B Semantic Similarity Threshold

This section explains the rationale behind the semantic similarity threshold of 0.7 used in the factual prediction experiment.

Language models with a relatively large number of parameters exhibit expressive capabilities. As a result, evaluating their outputs solely based on surface-level string matching is insufficient to capture their semantic appropriateness. To address this limitation, this study adopts a quantitative evaluation method based on semantic similarity.

For computing semantic similarity, this study employs all-MiniLM-L6-v2, a pre-trained language model based on Sentence-BERT (Reimers and Gurevych, 2019). Each word pair is encoded into embeddings, and cosine similarity is used to quantify their semantic closeness. Pairs with a similarity score of 0.7 or higher are considered semantically similar. This threshold is selected as a conservative and empirically grounded criterion to ensure a clearer separation between semantically similar and dissimilar cases, thereby enhancing consistency and precision in evaluation. Although lower similarity scores may sometimes correspond to semantically related expressions, a threshold of 0.7 is adopted to maintain reliability and interpretability.

To support the validity of this threshold, Table 4 presents examples of sentence pairs spanning a range of similarity scores. These examples demonstrate that semantic similarity–based evaluation enables more fine-grained and accurate assessment than methods relying solely on surface-level string matching.

Word A	Word B	Similarity Score
table	sadness	0.09
law	noodles	0.17
rocket	freedom	0.24
chair	depression	0.32
urinary	water	0.40
task	project	0.56
publicity	advertising	0.62
student	school	0.65
movie	cinema	0.68
combat	**fight**	**0.71**
computer	**laptop**	**0.71**
art	**painting**	**0.72**
middle	**mid**	**0.75**
city	**urban**	**0.86**
purchase	**buy**	**0.87**
bike	**bicycle**	**0.92**
sofa	**sofa**	**1.00**

Table 4: Examples of semantic similarity scores for word pairs measured by Sentence-BERT.

C Implementation Details

All language model families used in our experiments—including GPT, LLaMA, Qwen, and DeepSeek—are loaded via the Hugging Face Transformers[3] library, which provides a unified interface for model and tokenizer handling.

[3] https://huggingface.co/

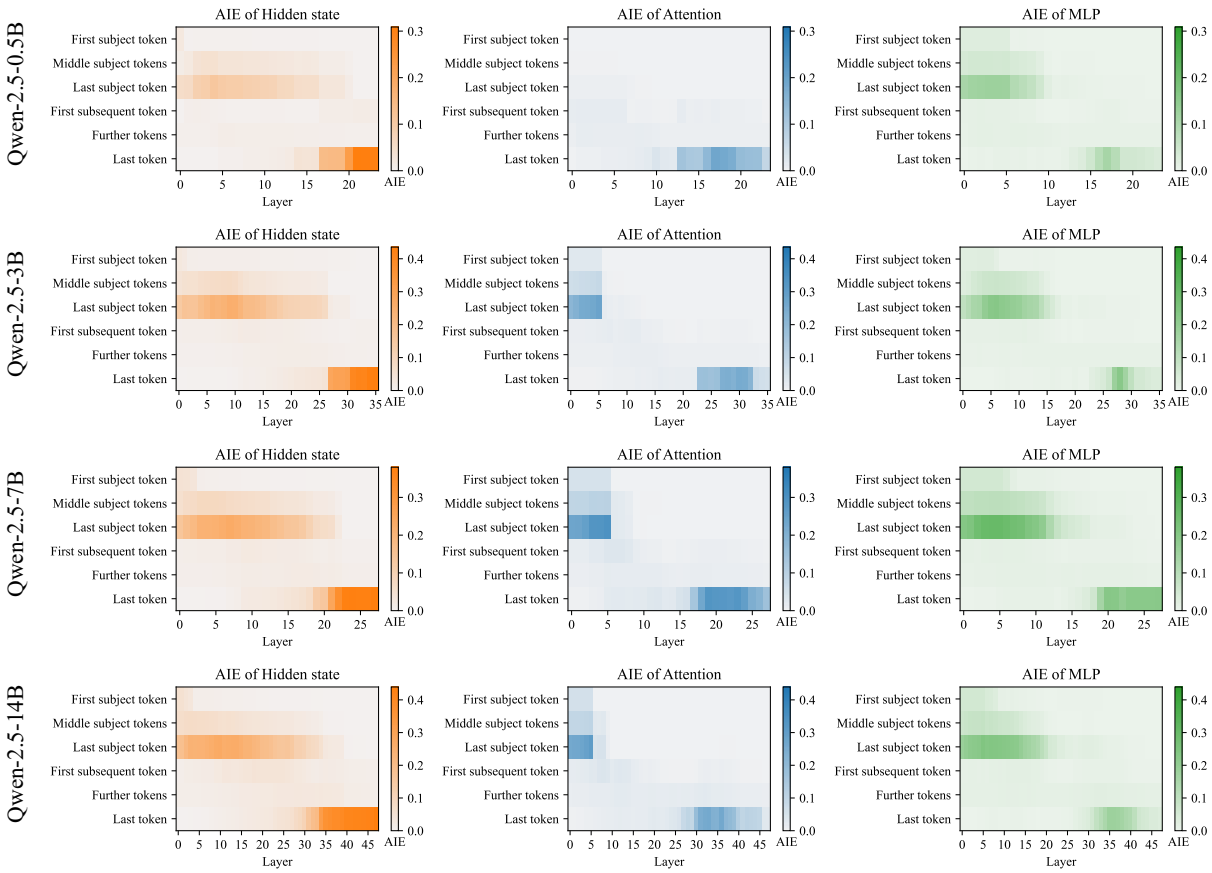

Figure 8: Restoration effects in Qwen-family autoregressive Transformer models with varying parameter scales.

During preprocessing as described in Section 4.3, we use NLTK[4] to remove stopwords from prompts. For candidate selection, we apply BM25 ranking using the rank_bm25[5] python package.

Meng et al. (2022a) is released under the MIT License, and Geva et al. (2023) under the Apache License 2.0, both of which we build upon and modify to support our evaluation framework and to extend their functionalities for a broader set of autoregressive Transformer architectures.

[4] https://www.nltk.org/
[5] https://github.com/dorianbrown/rank_bm25

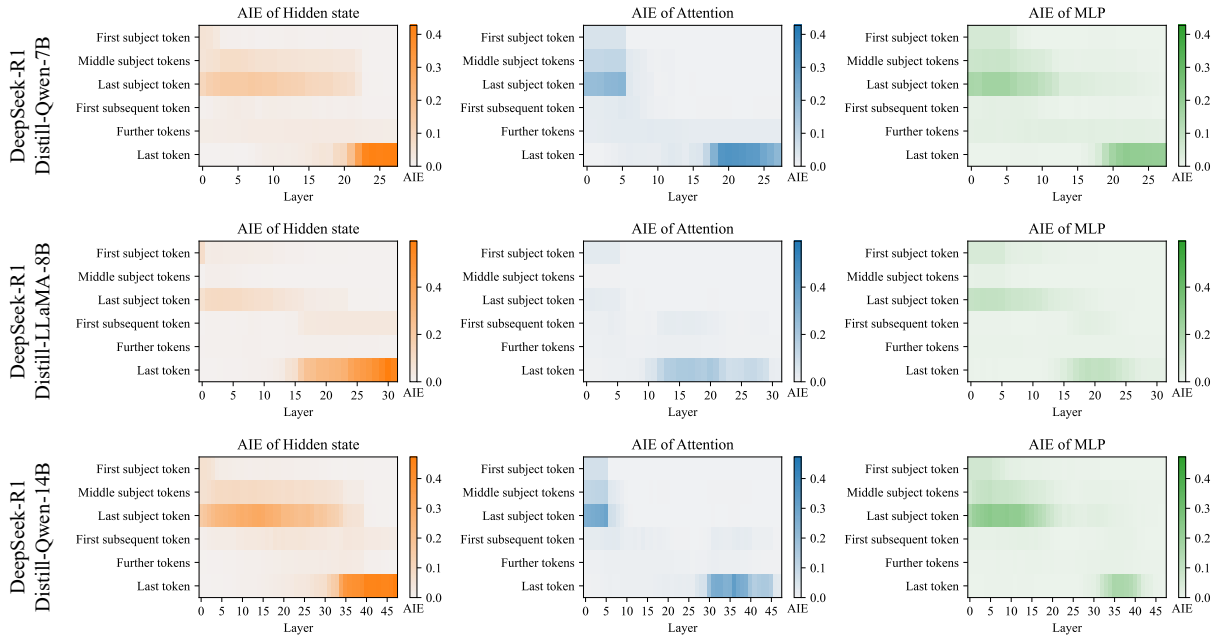

Figure 9: Restoration effects in DeepSeek-family autoregressive Transformer models with varying parameter scales.

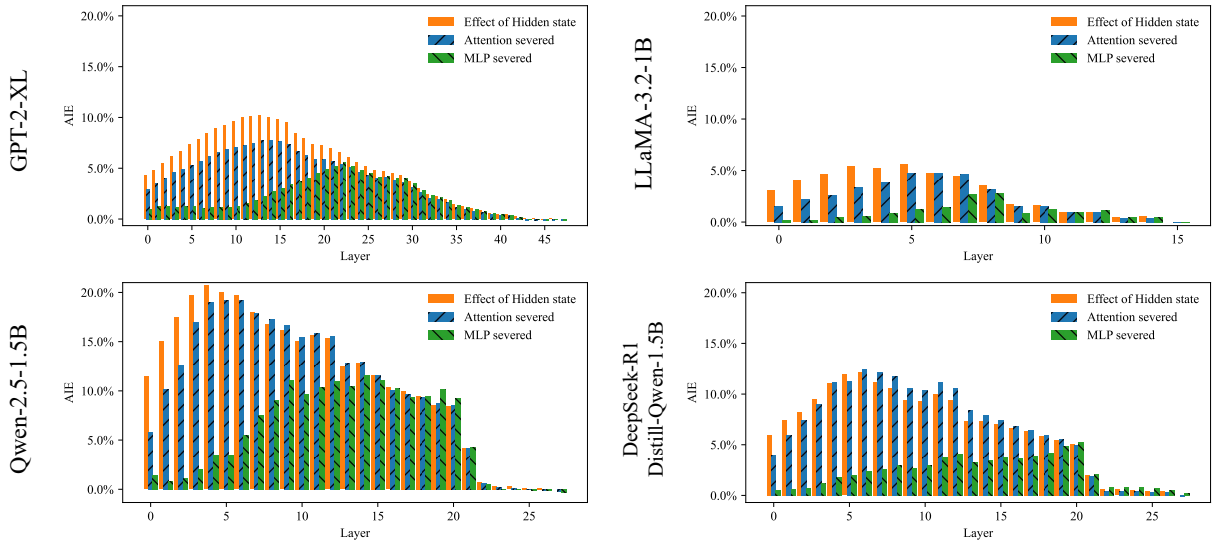

Figure 10: Severing effects across multiple autoregressive Transformer models (all layers).

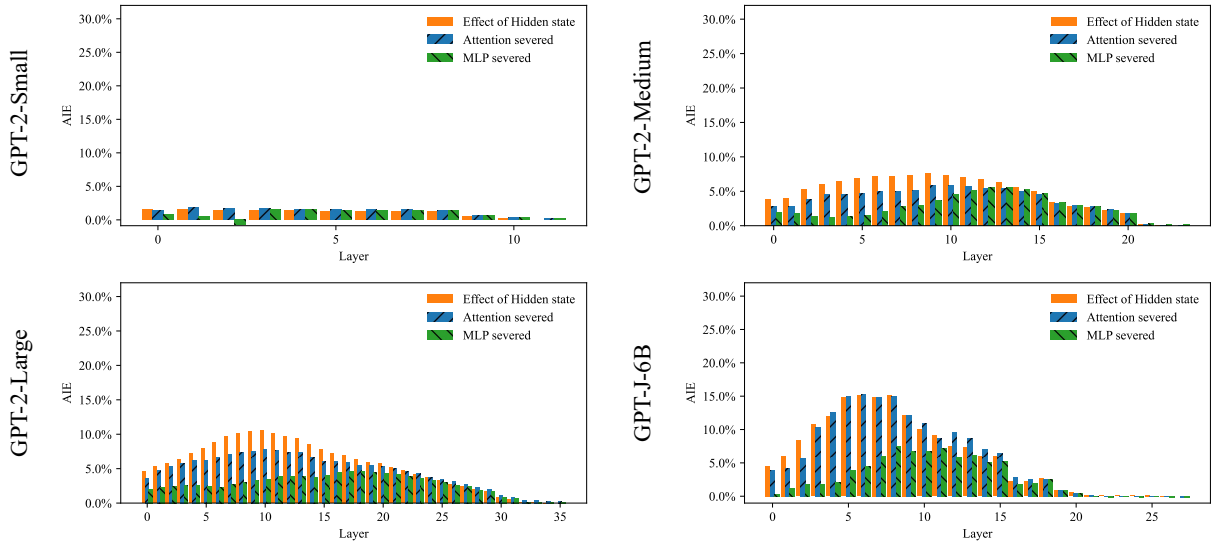

Figure 11: Severing effects across GPT-family autoregressive Transformer models at varying parameter scales.

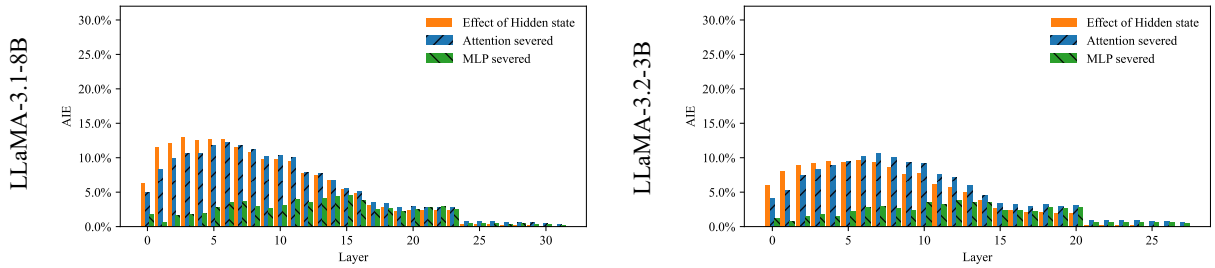

Figure 12: Severing effects across LLaMA-family autoregressive Transformer models at varying parameter scales.

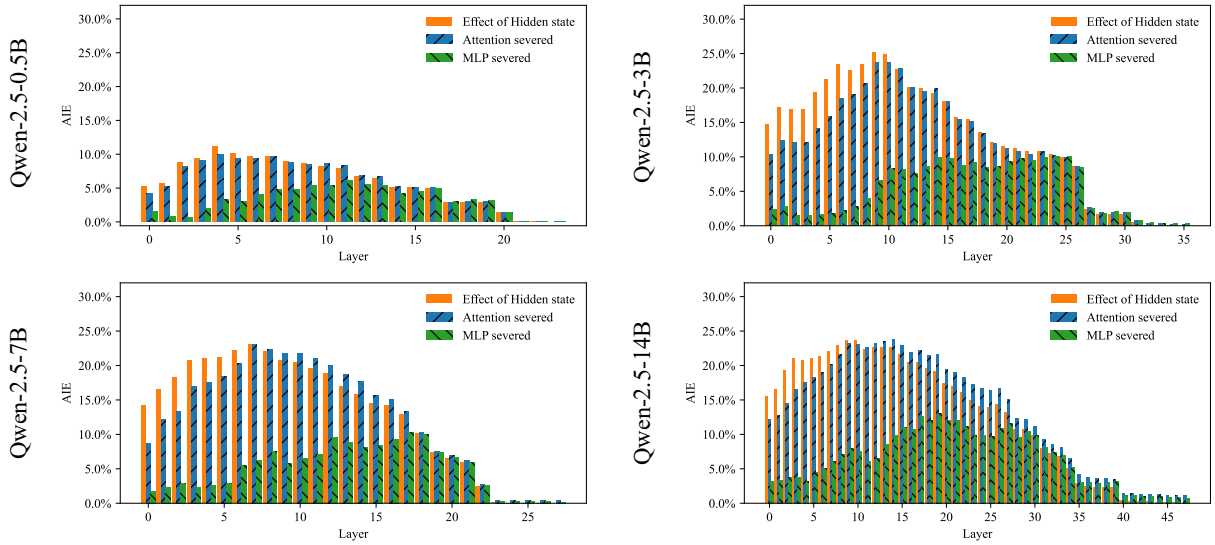

Figure 13: Severing effects across Qwen-family autoregressive Transformer models at varying parameter scales.

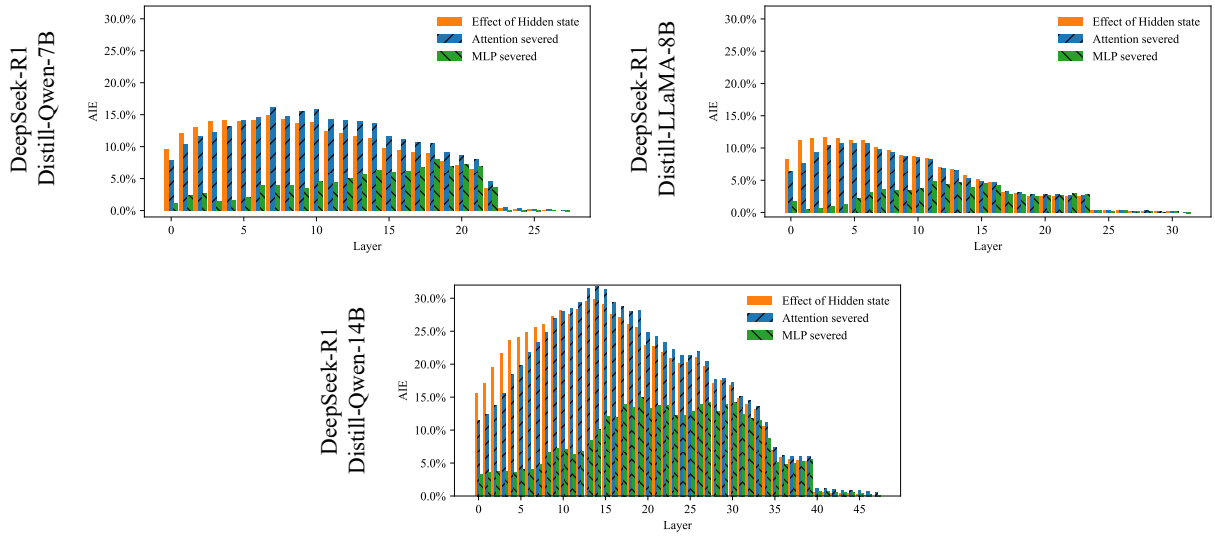

Figure 14: Severing effects across DeepSeek-family autoregressive Transformer models at varying parameter scales.

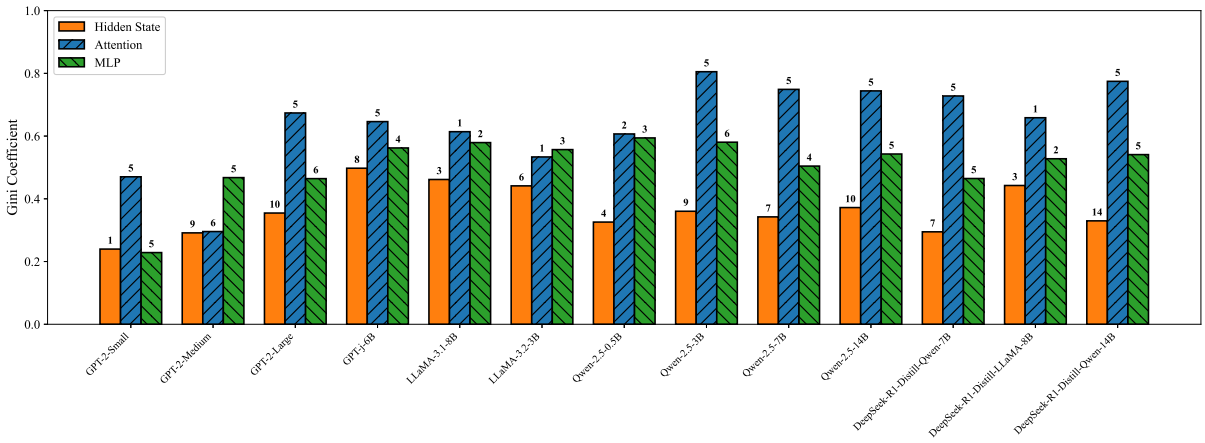

Figure 15: Gini coefficient-based concentration analysis of causal effects in Attention and MLP modules across autoregressive Transformer models with varying parameter scales (numbers above bars indicate the layer with the highest concentration).

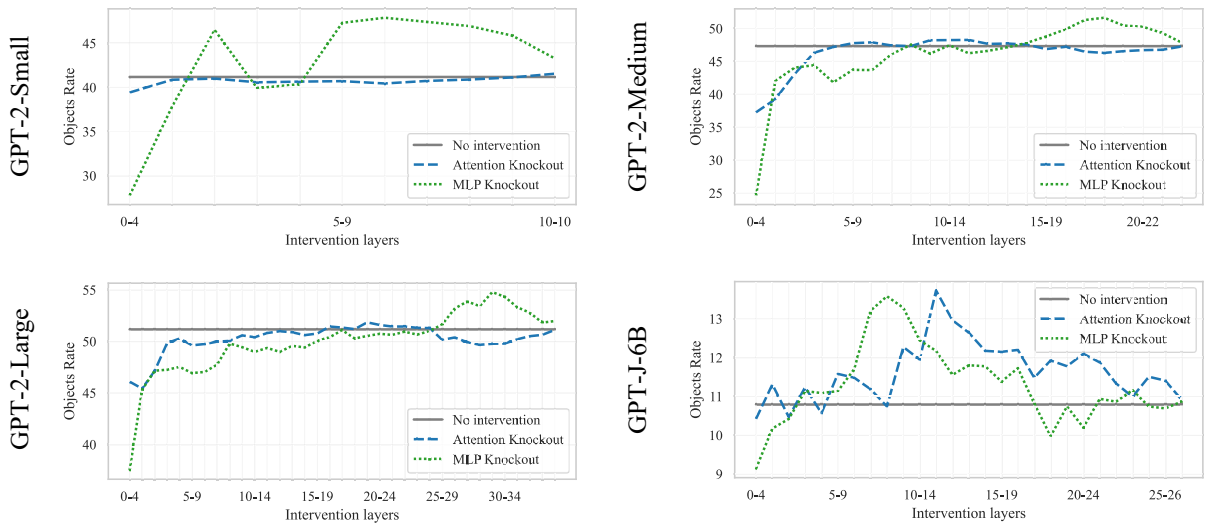

Figure 16: Factual prediction evaluation after knockout of Attention and MLP outputs across GPT-family autoregressive Transformer models with varying parameter scales.

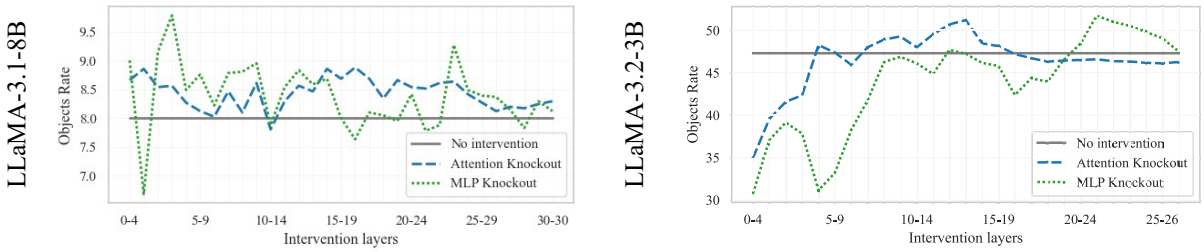

Figure 17: Factual prediction evaluation after knockout of Attention and MLP outputs across LLaMA-family autoregressive Transformer models with varying parameter scales.

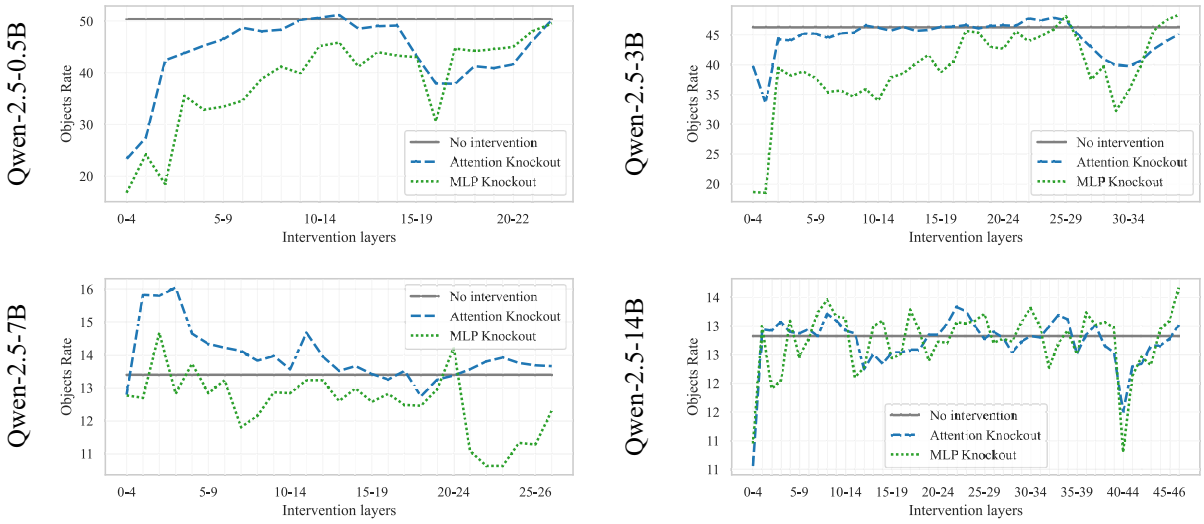

Figure 18: Factual prediction evaluation after knockout of Attention and MLP outputs across Qwen-family autoregressive Transformer models with varying parameter scales.

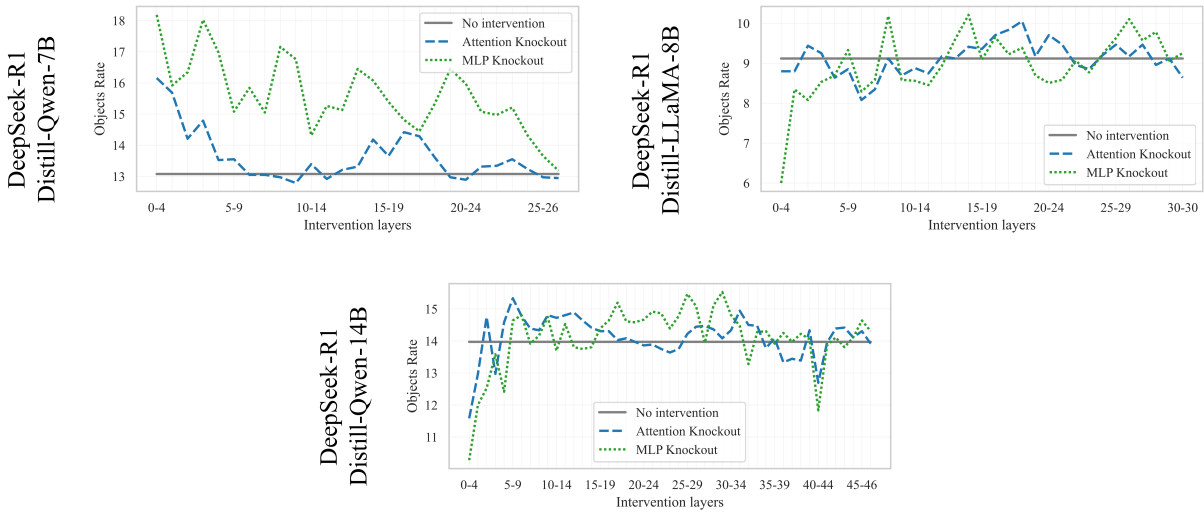

Figure 19: Factual prediction evaluation after knockout of Attention and MLP outputs across DeepSeek-family autoregressive Transformer models with varying parameter scales.