ART: Attention Replacement Technique to Improve Factuality in LLMs

Ziqin Luo[2]*, Yihao Quan[1]*, Xiaofeng Zhang[1]*†, Xiaosong Yuan[3], Chen Shen[3]
[1]Department of Automation and Intelligent Sensing, Shanghai Jiao Tong University
[2]Fudan University [3]Alibaba Cloud Computing
{framebreak@}sjtu.edu.cn

Abstract

Hallucination in large language models (LLMs) continues to be a significant issue, particularly in tasks like question answering, where models often generate plausible yet incorrect or irrelevant information. Although various methods have been proposed to mitigate hallucinations, the relationship between attention patterns and hallucinations has not been fully explored. In this paper, we analyze the distribution of attention scores across each layer and attention head of LLMs, revealing a common and intriguing phenomenon: Shallow layers of LLMs primarily rely on uniform attention patterns, where the model distributes its attention evenly across the entire sequence. This uniform attention pattern can lead to hallucinations, as the model fails to focus on the most relevant information. To mitigate this issue, we propose a training-free method called **A**ttention **R**eplacement **T**echnique (ART), which replaces these uniform attention patterns in the shallow layers with local attention patterns. This change directs the model to focus more on the relevant contexts, thus reducing hallucinations. Through extensive experiments, ART demonstrates significant reductions in hallucinations across multiple LLM architectures, proving its effectiveness and generalizability without requiring fine-tuning or additional training data.

1 Introduction

Large language models (LLMs) have made significant strides, but hallucinations remain a persistent challenge, particularly in tasks such as Question Answering (QA). This is a highly challenging issue for the practical deployment of large models. Therefore, to ensure that the model's output is true and reliable, existing research attempts to solve the hallucination problem by incorporating external knowledge bases or retraining models with additional data (Chen et al., 2024b). However, these approaches often require significant resources and may substantially reduce the inference efficiency.

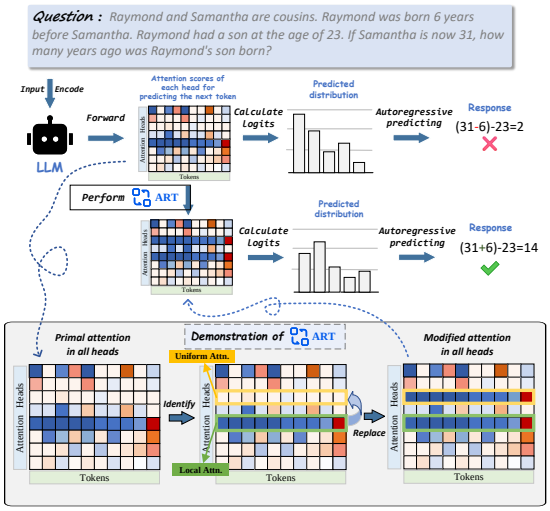

Figure 1: A demonstration overview of how ART works during the decoding process, with $N_h = 8$ and $k = 1$.

To strike a trade-off between performance and resource consumption, researchers have intervened in the attention mechanism of transformer-based large language models to mitigate hallucinations (Zhou et al., 2025; Huang et al., 2024; Zhang et al., 2024; Shi et al., 2025, 2026; Zhang and Zhang, 2025; Zhang et al., 2026b; Wei and Zhang, 2024a; Zhang et al., 2026a; Zhao et al., 2026; Zhang et al., 2025b,d). This empirical study provides valuable insights and reveals the important role that attention heads play in alleviating hallucinations. More broadly, recent studies have also examined consistency and reliability issues in adjacent multimodal and agent settings (Quan et al., 2026; Zhang et al., 2025a).

In addition to head-level analysis, there are also layer-level analyses. Recent research (Chen et al., 2024c; Zhang et al., 2025c; Che et al., 2026) reveals that shallow layers, particularly the first two layers, play a more critical role in knowledge in-

* These authors contributed equally to this work
† Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34068–34078

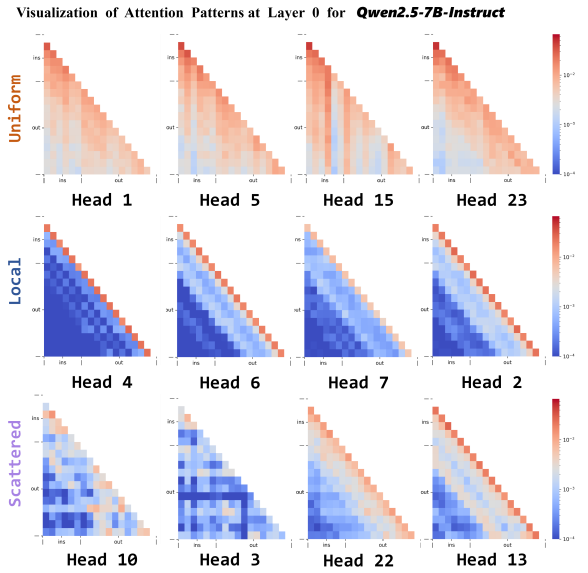

Figure 2: Visualization on some of attention weights of **Layer 0** in *Qwen2.5-7B-Instruct* model when encoding a sample from GSM8K.Attention weights are divided into 3 categories: *Uniform Attention*, *Local Attention*, and *Scattered Attention*.

jection and deserve denser injection, while deeper interventions have minimal effect and could even be pruned. AdaInfer (Fan et al., 2024) also supports this idea, suggesting that current LLMs can perform adequately if truncated at certain intermediate layers, as the middle and later layers do not contribute much more to the intermediate representations in some tasks.

Combining head-level and layer-level analyses and ideas, we believe that shallow-layer attention heads play an important role in mitigating hallucinations. Therefore, we visualize the attention maps of the various attention heads in shallow layers, as shown in Figure 2. These attention heads can be categorized into three types based on their distribution: local, uniform, and medium. The characteristics of these types are as follows: (1) **Uniform heads** distribute the model's attention evenly across the context. (2) **Local heads** only attends to neighboring tokens, and the visualized results are relatively sparse. (3) **Scattered heads** focuses predominantly on certain preceding tokens.

Since layers are not fully decoupled, a combination of most layers must be executed sequentially to achieve the best results (Sun et al., 2024). ITI (Li et al., 2023) also indicates that the local pattern is crucial for the model's understanding of semantics. Therefore, we aim to maximize the effectiveness of shallow layers in understanding semantics and contextual information, thus mitigating hallucinations while maintaining the integrity of the model.

Motivated by these observations, we propose an attention-enhancement method in shallow layers, which replaces uniform attention heads with local attention heads to enhance the information exchange capability of the model's attention, and thus improve the model's performance. We conduct extensive evaluations, specifically focusing on hallucination issues, and test mainstream LLMs to validate the effectiveness of ART in reducing hallucinations across various model architectures. Our results demonstrate that ART is a highly effective plug-and-play solution for mitigating hallucinations in various LLMs.

Specifically, our contributions can be summarized as follows:

- We systematically studied the distribution characteristics of shallow-layer attention heads, categorizing them into three types: Uniform, Local, and Scattered. Through experiments, we validated the crucial role of local attention in the model's generation process.
- To address the hallucination problem in large language models (LLMs), we propose a training-free method called Attention Replacement Technique (ART) to replaces redundant uniform attention heads with local attention heads in the shallow layers, significantly improving the truthfulness of the model's output.
- Experiments on multiple models validate the plug-and-play convenience and strong generalization of this method.

2 Related Work

Attention sink and information flow. StreamingLLM (Xiao et al., 2024b) observes high attention values on the first token, termed an "attention sink," and leverages this finding to extend the input sequence length—a positive use of high attention values. However, different scenarios can exhibit different behaviors. Several studies have demonstrated the negative effects of the attention sink phenomenon. The ACT (Yu et al., 2024) study found that attention sinks not only occur on the first token but also on tokens with limited semantic information (e.g., ".", ":", and "<0x0A>") . Contrary to the observations made in StreamingLLM—which suggest preserving attention sinks to enhance LLMs' accuracy—they highlight that not all attention sinks are beneficial.

Specifically, for most attention sinks occurring in the middle or later parts of inputs, reducing their attention scores can lead to improved accuracy. In OPERA (Huang et al., 2024) and DOPRA (Wei and Zhang, 2024b), it was found that when models generate hallucinated content, the self-attention weights in the last layer exhibit a distinct "columnar" pattern before the hallucination occurs, leading to an "over-trust" tendency in the self-attention weights for the hallucinated parts. Yuan et al. (2024) proposed Instance-adaptive prompting to select better prompt to LLMs for correct reasoning for different instances, and IAP-ss compared the significance scores and thresholds of the information flow to determine the appropriate prompt. EAH (Zhang et al., 2024) reveals a phenomenon: most hallucinations are closely related to the attentional sinking pattern in the image-labeled self-attention matrix, where shallower layers show intensive attentional sinking and deeper layers show sparse attentional receptions. They propose a training-free approach called Enhanced Attention Head (EAH) designed to enhance image convergence by focusing attention on shallow layers.

Attention Heads of Large Language Models. LoFiT (Yin et al., 2024)'s extends the task from model realism more generally, for a particular downstream task, the first step is to first identify the subset of attentional heads that are most important for the task by learning, and then the second step is to train a corresponding offset vectors for each of the attentional heads in this subset of heads, with the core purpose is to optimize the activations of these heads. Michel et al. (2019) argues that the nlp model, even if trained in the mode of multi-head attention, can remove a large fraction of these attention heads during real-world testing without affecting the model's performance. Wu et al. (2024) argues that all models with long context capabilities have a set of retrieval heads, and pruning out the retrieval heads will result in the loss of the model's ability to retrieve relevant information and create illusions, whereas pruning the non-retrieval heads does not affect the model's retrieval ability. Xiao et al. (2024a) argues that the attention heads within the LLM can be roughly divided into two parts: retrieval heads and streaming heads; the former is important for the model to find key information in context, and cropping the (contextual) KV cache within retrieval heads significantly affects the final output of the language model, while modifying the KV cache of streaming heads has no The final output of the language model is significantly affected by cropping the (contextual) KV cache within retrieval heads while modifying the KV cache of streaming heads does not.

3 Preliminaries

A decoder-only LLM is composed of L stacked decoder layers. Each layer consists of two modules: Multi-Head Attention (MHA) and Feed-Forward Network (FFN). They are connected serially and encode the T input tokens $\boldsymbol{t} = \{t_1, t_2, \ldots, t_T\}$ jointly. Formally, Let the input of the l-th layer be $\boldsymbol{H}^l = \{\boldsymbol{h}_1^l, \boldsymbol{h}_2^l, \ldots, \boldsymbol{h}_T^l\} \in \mathbb{R}^{T\times d}$. The t-th token is encoded with a densen hidden state $\boldsymbol{h}_t^l$ with dimension being d. Since LLMs mainly adopt a pre-norm structure, we represent the l-th layer:

$$\boldsymbol{O}^l = \text{MHA}^l\left(\text{LN}(\boldsymbol{H}^l)\right) \quad (1)$$

$$\boldsymbol{H}^{l+1} = \text{FFN}^l(\text{LN}(\boldsymbol{O}^l + \boldsymbol{H}^l)) + \boldsymbol{O}^l + \boldsymbol{H}^l. \quad (2)$$

MHA^l is the l-th layer's multi-head attention operator, projecting the input sequence $\boldsymbol{X} \in \mathbb{R}^{T\times d}$ into N_h subspaces to perform attention mechanisms in parallel and aggregating them at last:

$$\text{MHA}^l(\boldsymbol{X}) = \sum_{h=1}^{N_h} \text{Softmax}\left(\frac{\boldsymbol{Q}_h^l(\boldsymbol{K}_h^l)^\top}{\sqrt{d_h}}\right)\boldsymbol{V}_h^l\boldsymbol{W}_{o,h}^l \quad (3)$$

$$\boldsymbol{Q}_h^l = \boldsymbol{X}\boldsymbol{W}_{q,h}^l, \boldsymbol{K}_h^l = \boldsymbol{X}\boldsymbol{W}_{k,h}^l, \boldsymbol{V}_h^l = \boldsymbol{X}\boldsymbol{W}_{v,h}^l. \quad (4)$$

Here, $\boldsymbol{W}_{q,h}^l, \boldsymbol{W}_{k,h}^l, \boldsymbol{W}_{v,h}^l \in \mathbb{R}^{d\times d_h}$ and $\boldsymbol{W}_{o,h}^l \in \mathbb{R}^{d_h\times d}$ are parametric matrices belonging to the h-th head of l-th layer. $d_h = d/N_h$ represents the dimension of each projected head. To investigate how each token in the sequence attends itself and its previous tokens, we denote the attention weight of the h-th head as

$$\boldsymbol{A}_h^l \equiv \text{Softmax}\left(\frac{\boldsymbol{Q}_h^l(\boldsymbol{K}_h^l)^\top}{\sqrt{d_h}}\right). \quad (5)$$

Actually, attention weights vary among different heads, capturing different patterns in the input sequences (Vaswani et al., 2017). Thus, MHA makes language models more expressive than vanilla attention. In the remainder of this paper, unless otherwise specified, the concept of *attention* refers to the attention weight $\boldsymbol{A}_h^l$.

3.1 Attention Patterns in LLMs

Recent studies (Zheng et al., 2024; Xiao et al., 2024b; Yu et al., 2024; Chen et al., 2024a) have

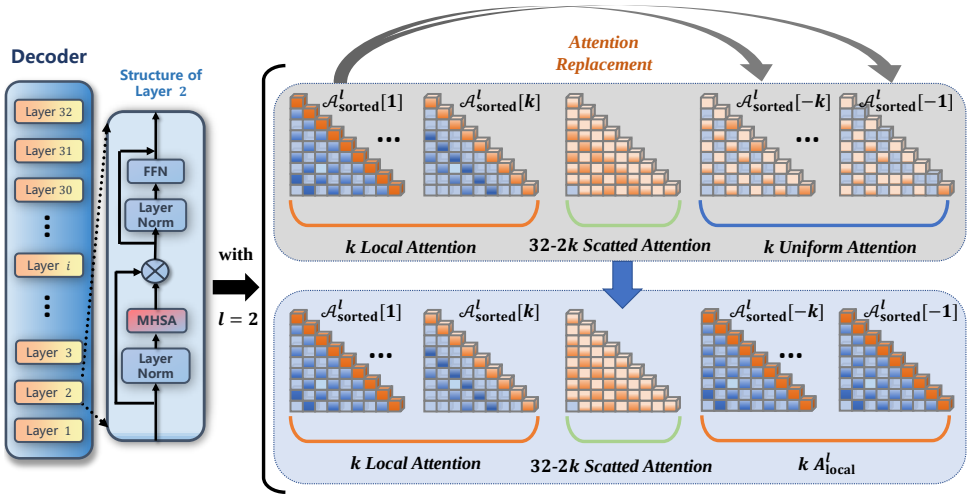

Figure 3: The detailed demonstration of applying ART-max to the second layer of *Llama2-7B-Chat*.

investigated the internal attention heads of LLMs as they process input sequences, providing detailed visualizations of attentions across different layers. These studies reveal that attentions in LLMs exhibit diverse patterns. We aim to investigate these different attention patterns and explore their influences on LLMs' generation. To achieve this goal, we categorize attention patterns (Section 3.2) and examine their effects (Section 3.3).

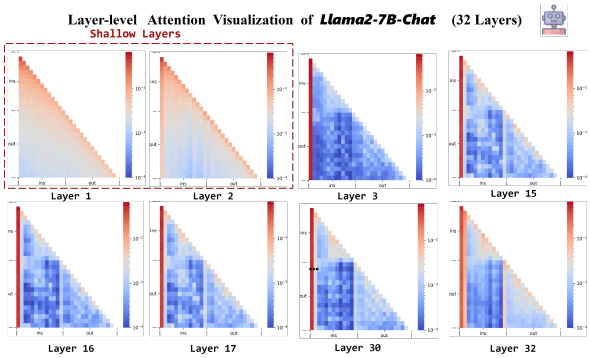

Figure 4: The structure diagram of ART replaces the uniform attention head in each attention head that predicts the next token with a local attention head. Baseline: *Llama2-7B-Chat* processes a sample from TruthfulQA.

3.2 Category

Sparse and dense attention. Yu et al. (Yu et al., 2024) and Chen et al. (Chen et al., 2024a) visualize multiple attentions across various layers within LLMs. We can categorize attention patterns into *sparse attention* and *dense attention* based on the presence of attention sinks (Xiao et al., 2024b). Specifically, attention sinks are tokens that absorb a significant amount of attention scores from other tokens that can attend them, resulting in sparse attention patterns where the attention scores obtained by other tokens are dramatically low. In contrast, when attention sinks are absent, the attention scores are more evenly distributed among tokens, corresponding to dense attention. As shown in Figure 4, visually, in sparse attention, most tokens receive nearly zero attention, resulting in a generally darker appearance. In dense attention, the attention distribution is relatively balanced, leading to a lighter overall color. Generally, sparse attention predominantly occurs in LLMs' deeper layers, while dense attention appears primarily in the shallower layers. In particular, Chen et al. (Chen et al., 2024b) demonstrate that shallow layers are more important for the injection and understanding of knowledge compared to deeper layers. FastV's experiments (Chen et al., 2024a) also indicate that editing shallow layers has a greater impact on the generation process than editing deep layers. Therefore, this paper discusses attentions in the shallow layers, aiming to intervene in attention as minimally as possible. Following the settings of previous studies (Chen et al., 2024a; Yu et al., 2024), unless otherwise specified, we define shallow layers as the first two layers of LLMs.

Three patterns for dense attention. Different attention heads exhibit different attention patterns in shallow layers of dense attention. As illustrated in Figure 2, dense attention can generally be categorized into three patterns: uniform, local, and scattered. These represent three different characteristics of attention distribution. In uniform attention, each token attends almost equally to preceding tokens; in local attention, each token only attends

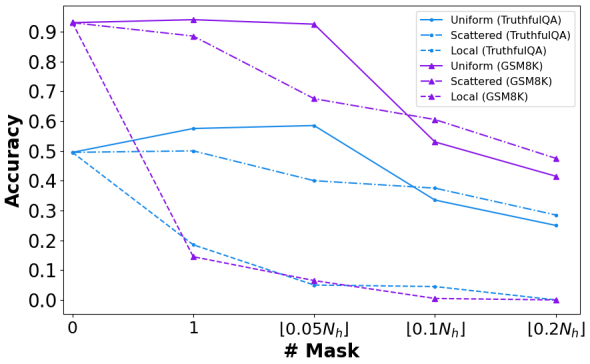

Figure 5: Accuracy curves of *Qwen2.5-7B-Instruct* evaluated on $\mathcal{D}_{\text{ablation}}$ of TruthfulQA and GSM8K subsets after masking different amounts of *uniform*, *scattered*, and *local* attention.

to neighboring tokens; in scattered attention, each token focuses predominantly on certain preceding tokens. For better illustration, we define the m-index m_h^l to classify these three types of attention. Formally, let $\boldsymbol{U} \in \mathbb{R}^{T\times T}$ denotes the completely uniform attention

$$\boldsymbol{U}[i,j] = \begin{cases} 0 & i < j \\ 1/i & i \geq j \end{cases}, \tag{6}$$

where $\boldsymbol{U}[i,j]$ is the attention weight between the i-th and j-th tokens in input tokens $\boldsymbol{t}$. We define $\boldsymbol{A}_h^l$ as the attention of the h-th head in the l-th layer and m_h^l as:

$$m_h^l = \frac{1}{N(N+1)/2} \sum_{i\geq j} \max\left(\frac{\boldsymbol{A}_h^l[i,j]}{\boldsymbol{U}_h^l[i,j]}, \frac{\boldsymbol{U}_h^l[i,j]}{\boldsymbol{A}_h^l[i,j]}\right) \in \mathbb{R}^+. \tag{7}$$

The smaller m_h^l, the more similar $\boldsymbol{A}_h^l$ is to $\boldsymbol{U}_h^l$, corresponding to uniform attention. In contrast, a larger value of m_h^l signifies a greater distinction between $\boldsymbol{A}_h^l$ and $\boldsymbol{U}_h^l$, which refers to local attention. The transition from uniform attention to local attention can be viewed as a continuous spectrum characterized by their corresponding m_h^l. Scattered attention represents an intermediate state along this spectrum. Since the absolute value of the m-index is influenced by the sequence length T, the relative magnitude of different m-indices is more meaningful. Specifically, for all attention heads within the same layer, we calculate their respective m-indices and rank these indices by magnitude. Attentions ranked at the top are considered uniform attention, while those at the bottom are regarded as local attention. The rest in the middle of the ranking are considered scattered attention.

3.3 Influence on Model Generation

Xiao et al. (Xiao et al., 2024b) and Yu et al. (Yu et al., 2024) studied the role of attention sink in sparse attention. However, the influence of different attention patterns in dense attention is not yet clear. Therefore, we investigate the effect of the three attention patterns on LLMs' accuracy in downstream tasks. We conduct preliminary experiments on four datasets that cover diverse scenarios. Specifically, we randomly sample 200 data points each from the TruthfulQA (Lin et al., 2022), GSM8K (Cobbe et al., 2021), LogiQA (Liu et al., 2020), and CommonsenseQA (Talmor et al., 2019) datasets (denoted as $\mathcal{D}_{\text{ablation}}$ in Section 4.5) and test them using *Qwen2.5-7B-Instruct*. We record the changes in the model's accuracy when different proportions of uniform, local, and dispersed attention are masked. From Figure 5, we derive three observations: i) The local attention is crucial for model generation. Masking even a single local attention significantly impairs the model generation; ii) Uniform and scattered attention exhibit a degree of redundancy. Masking small amounts of these patterns has few influence on model generation, and in some cases, can enhance it. iii) When a small enough amount of attention is masked, uniform attention shows a higher redundancy than scattered attention, thus influencing the model generation less when masked. Based on these observations, we consider leveraging the redundancy present in LLMs' dense attention by replacing the most redundant uniform attentions with local attentions to improve attention utilization during generation, thereby enhancing LLMs' performance.

3.4 Attention Replacement Technique

In light of our findings, we propose a lightweight and effective Attention Replacement Technique (ART) to enhance LLMs by intervening the multi-head attention module during the model generation. Assuming there are a total of N_h attention heads, for the multi-head attention module MHA^l in the l-th layer, we can rewrite it as

$$\text{MHA}^l(\boldsymbol{X}) = \sum_{h=1}^{N_h} \boldsymbol{A}_h^l f_h^l(\boldsymbol{X}). \tag{8}$$

Let $\mathcal{A}^l = \{\boldsymbol{A}_1^l, \boldsymbol{A}_2^l, \ldots, \boldsymbol{A}_{N_h}^l\}$ represent all the attentions. According to equation (7), we can compute the corresponding metrics $\boldsymbol{m}^l = \{m_1^l, m_2^l, \ldots, m_{N_h}^l\}$. By sorting $\mathcal{A}^l$ in ascending

order based on $\boldsymbol{m}^l$, we obtain $\mathcal{A}^l_{\text{sorted}}$. We define k to denote that we consider the first k attention heads with the smallest m-indices in $\mathcal{A}^l_{\text{sorted}}[:k]$ as uniform attentions and the last k attention heads with the largest m-indices in $\mathcal{A}^l_{\text{sorted}}[-k:]$ as local attentions. We then define $\boldsymbol{A}^l_{\text{local}}$ as

$$\textit{Max: } \boldsymbol{A}^l_{\text{local}} = \mathcal{A}^l_{\text{sorted}}[-1] \quad \text{or} \tag{9}$$

$$\textit{Mean: } \boldsymbol{A}^l_{\text{local}} = \frac{1}{k}\sum_{j=1}^{k} \mathcal{A}^l_{\text{sorted}}[-j]. \tag{10}$$

Max indicates the local attention is selected with the maximum m-index as $\boldsymbol{A}^l_{\text{local}}$, whereas *Mean* calculates $\boldsymbol{A}^l_{\text{local}}$ by averaging all local attentions. If $\boldsymbol{A}^l_{\text{local}}$ is calculated through equation (9), ART is specified as ART-max. Otherwise, ART is denoted as ART-mean. We replace $\mathcal{A}^l_{\text{sorted}}[:k]$ with $\boldsymbol{A}^l_{\text{local}}$ to improve LLMs' attention utilization. Therefore, equation (8) can be rewritten as

$$\begin{aligned}\text{ART-MHA}^l(\boldsymbol{X}) = &\sum_{i\in\mathcal{I}} \boldsymbol{A}^l_{\text{local}} f^l_i(\boldsymbol{X}) \\ &+ \sum_{i\notin\mathcal{I}} \boldsymbol{A}^l_i f^l_i(\boldsymbol{X}),\end{aligned} \tag{11}$$

where $\mathcal{I}$ is the index set corresponding to $\mathcal{A}^l_{\text{sorted}}[:k]$. The operation ART-MHA^l is then applied to model inference for efficient intervention. The full pipeline is demonstrated in Algorithm 1.

4 Experiment

4.1 Datasets and Models

Datasets. We evaluate ART on three datasets: (i) TruthfulQA (Lin et al., 2022) measures whether LLMs are truthful in responding to questions. (ii) LogiQA (Liu et al., 2020), consisting of expert-written questions, aims to test logical reasoning. (iii) GSM8K (Cobbe et al., 2021), containing high-quality elementary school math word problems, is targeted at measuring LLMs' mathematical abilities. These datasets cover multiple scenarios for the application of LLMs and are suitable for comprehensive evaluation of their capabilities.

Models. ART is evaluated on prevalent open-source LLMs. They are Llama2-7B-Chat (Touvron et al., 2023), Llama3.1-8B-Instruct (AI@Meta, 2024), Ministral-8B-Instruct-2410[1], Qwen2-7B-Instruct (Yang et al., 2024), and Qwen2.5-7B/14B/32B-Instruct (Yang et al., 2024).

[1]Ministral blog post.

4.2 Baselines and Metrics

Following a similar setup (Li et al., 2023; Yu et al., 2024), we compare ART with the vanilla decoding baseline under the zero-shot Chain-of-Thought (Wei et al., 2022) setting. We utilize accuracy as the metric for the evaluation of performance. Notably, questions from GSM8K ask models to answer without answer options available, while questions from other datasets are presented as Multi-Choice Questions (MCQs).

4.3 Implementation Details

We run all our experiments on 1 NVIDIA A800 GPU with 80GB memory. For TruthfulQA, we directly adopt the whole test dataset for evaluation. For other datasets, we uniformly sample 1,000 from each to build the test datasets $\mathcal{D}_{\text{test}}$, thus balancing the experimental cost and effectiveness. All experiments are conducted in a zero-shot CoT setting. Detailed prompt templates are available in our repository. In all our experiments, unless otherwise specified, we use $k = \lfloor 0.1 * N_h \rfloor$, which is derived through the discussion in Section 4.5.

4.4 Main Results

Table 1 presents the experimental results that we validate ART on previously mentioned datasets that cover multiple LLM application scenarios. The effects of ART on different models across various task types differ. In general, ART can improve the accuracy of LLMs by 0.6% to 1.7% in most cases, and in some cases it can achieve performance improvements exceeding 3%. For example, applying ART-mean to *Qwen2-7B-Instruct* results in a 3.2% improvement in LogiQA. Similarly, applying ART-max to *Qwen2.5-7B-Instruct* achieves a 3.4% improvement in TruthfulQA. In terms of task types, compared to mathematical reasoning, ART exhibits better enhancement in model truthfulness and logical reasoning. The overall improvement for the GSM8K task is averaged 0.7%. Furthermore, for TruthfulQA and LogiQA, ART generally provides over a 1.1% improvement for LLMs. Notably, ART neither requires fine-tuning nor utilises in-domain data to determine hyperparameters or construct auxiliary models. Therefore, the performance gains ART provides to LLMs are quite substantial. Regarding model size, ART offers more accuracy enhancement to 7B/8B models than 14B/32B models. ART enhances *Qwen2.5-14B/32B-Instruct* models' accu-

Model	Method	Dataset Performance (%)					
		TruthfulQA	LogiQA	CommonsenseQA	OpenBookQA	GSM8K	Avg.
Llama2-7B-Chat	Vanilla	19.0	28.4	50.8	45.9	27.2	34.3
	ART-max	21.7 (+2.7)	29.7 (+1.3)	49.5 (-1.3)	44.9 (-1.0)	28.3 (+1.1)	34.8 (+0.5)
	ART-mean	21.5 (+2.5)	30.1 (+1.7)	49.6 (-1.2)	45.2 (-0.7)	28.0 (+0.8)	34.9 (+0.6)
Llama3.1-8B-Instruct	Vanilla	46.5	38.8	75.5	84.4	88.1	66.7
	ART-max	46.3 (-0.2)	39.6 (+0.8)	76.4 (+0.9)	85.4 (+1.0)	89.2 (+1.1)	67.4 (+0.7)
	ART-mean	46.8 (+0.3)	40.2 (+1.4)	76.0 (+0.5)	85.2 (+0.8)	89.0 (+0.9)	67.4 (+0.7)
Ministral-8B-Instruct-2410	Vanilla	46.1	41.1	71.4	84.3	89.8	66.5
	ART-max	46.4 (+0.3)	44.0 (+2.9)	73.6 (+2.2)	84.1 (-0.2)	90.2 (+0.4)	67.6 (+1.1)
	ART-mean	46.2 (+0.1)	43.5 (+2.4)	74.3 (+2.9)	84.3 (+0.0)	90.0 (+0.2)	67.7 (+1.2)
Qwen2-7B-Instruct	Vanilla	41.7	42.3	73.6	85.1	88.0	66.1
	ART-max	43.2 (+1.5)	45.0 (+2.7)	73.5 (-0.1)	84.6 (-0.5)	88.9 (+0.9)	67.0 (+0.9)
	ART-mean	43.8 (+2.1)	45.5 (+3.2)	73.3 (-0.3)	84.8 (-0.3)	88.8 (+0.8)	67.2 (+1.1)
Qwen2.5-7B-Instruct	Vanilla	51.2	50.7	78.1	87.2	93.2	72.1
	ART-max	54.6 (+3.4)	52.1 (+1.4)	79.6 (+1.5)	87.6 (+0.4)	93.8 (+0.6)	73.5 (+1.4)
	ART-mean	54.0 (+2.8)	52.5 (+1.8)	79.2 (+1.1)	87.2 (+0.0)	93.5 (+0.3)	73.3 (+1.2)
Qwen2.5-14B-Instruct	Vanilla	62.3	56.9	81.5	93.0	94.3	77.6
	ART-max	63.1 (+0.8)	57.8 (+0.9)	82.4 (+0.9)	92.8 (-0.2)	94.8 (+0.5)	78.1 (+0.5)
	ART-mean	63.3 (+1.0)	57.6 (+0.7)	82.0 (+0.5)	93.0 (+0.0)	94.8 (+0.5)	78.1 (+0.5)
Qwen2.5-32B-Instruct	Vanilla	71.6	63.0	84.6	95.0	94.7	81.8
	ART-max	72.0 (+0.4)	63.2 (+0.2)	84.9 (+0.3)	95.2 (+0.2)	95.4 (+0.7)	82.1 (+0.3)
	ART-mean	71.8 (+0.2)	63.5 (+0.5)	84.8 (+0.2)	95.1 (+0.1)	95.5 (+0.8)	82.1 (+0.3)

Table 1: Extended performance comparison between ART and vanilla decoding across six reasoning benchmarks. Experiments are conducted over $\mathcal{D}_{\text{test}}$.

racy in three tasks by approximately 0.7%, while it improves *Ministral-8B-Instruct-2410*, *Qwen2-7B-Instruct*, and *Qwen2.5-7B-Instruct* models by more than 1.1%. For the ART variants, the ART-max and ART-mean operations exhibit some differences in their calculation of $\boldsymbol{A}^{l}_{\text{local}}$. However, the experimental results indicate that their effect on model generation is almost identical. In practice, we can choose the appropriate operation in a flexible way. In subsequent experiments, unless otherwise specified, we default to implementing ART as ART-mean. Table 1 shows that ART provides robust improvement gains in LLMs, suggesting that ART can be applied in multiple tasks to enhance LLMs.

4.5 Ablation Studies

Practicing ART involves factors such as determining hyperparameter k and choice of attention heads, etc. To explore the effect of these factors on ART, we randomly sample 200 examples from the training set of each dataset to construct $\mathcal{D}_{\text{ablation}}$. We test ART in different settings on $\mathcal{D}_{\text{ablation}}$ to better understand and utilize it.

***k* selection**. As mentioned in Section 4.3, we set $k = \lfloor 0.1 * N_h \rfloor$ for our experiments. In fact, the selection of k significantly impacts the effectiveness of ART. Therefore, we investigate the variation in the accuracy of LLMs in $\mathcal{D}_{\text{ablation}}$ when k takes different values from the set $\{0, 1, \lfloor 5\% * N_h \rfloor, \lfloor 10\% * N_h \rfloor, \lfloor 20\% * N_h \rfloor, \lfloor 30\% * N_h \rfloor\}$. The results, as shown in Figure 6, indicate that in most cases, as k increases, the accuracy of the model increases first and then decreases. When $k = \lfloor 0.1 * N_h \rfloor$, the accuracy reaches the maximum. Beyond this point, further increasing k leads to more uniform attentions being identified and replaced, which can overly impair the model's context understanding ability, thereby resulting in a significant accuracy decrease. Consequently, we empirically set $k = \lfloor 0.1 * N_h \rfloor$ to achieve the optimal effectiveness of ART.

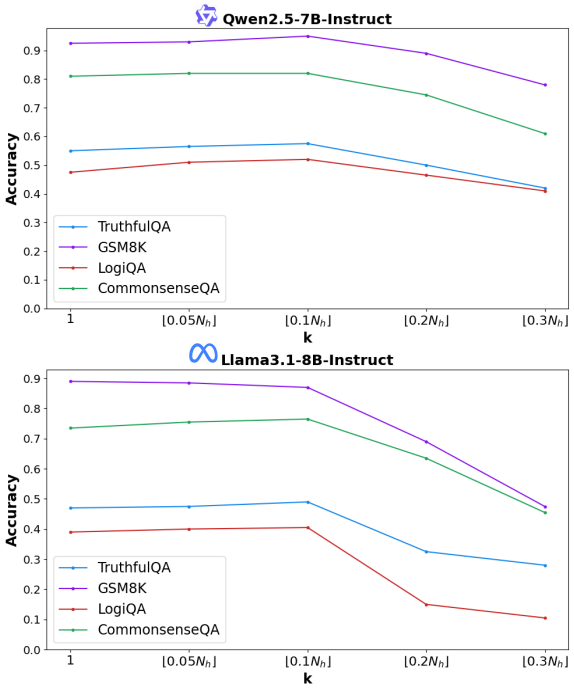

Figure 6: Accuracy curves of *Qwen2.5-7B-Instruct* and *Llama3.1-8B-Instruct* evaluated on $\mathcal{D}_{\text{ablation}}$ over different k values.

LLM	Method	TruthfulQA	LogiQA	GSM8K
Qwen2.5-7B-Inst	ART	52.5	51.0	94.0
	Vanilla	51.2	50.7	93.2
	ART-inverse	0.0	2.0	7.5
	ART-scattered	53.2	51.7	93.5
Llama3.1-8B-Inst	ART	46.0	40.5	87.0
	Vanilla	46.5	38.8	88.1
	ART-inverse	38.5	31.5	64.5
	ART-scattered	46.6	39.3	89.2

Table 2: Performance comparison between ART and ART-inverse. Tests are conducted on $\mathcal{D}_{\text{ablation}}$.

Replacement direction. Section 3.3. As discussed in Section 3.3, local attention is crucial for model generation, and masking local attention could make LLMs fail to complete tasks. An additional question arises: what if local attention is not masked but replaced with uniform attention? To investigate this, we test *ART-inverse*, which reverses the substitution in ART by replacing local attention with uniform attention. Table 2 conveys that for *Qwen2.5-7B-Instruct*, ART renders it almost entirely ineffective in downstream tasks, while for *Llama3.1-8B-Instruct*, ART leads to a significant performance degradation. Results from Table 1 and Table 2 collectively demonstrate that local attention is crucial, and modifications to local attention could severely impair model performances.

Ways to replace attentions. In Section 3.4, we describe ART in detail. Here, $\boldsymbol{A}^{l}_{\text{local}}$ the target attention, is calculated through local attention. A direct question is whether scattered attention can also be utilized to compute the target attention to enhance model generation. Therefore, we refer to this method as ART-scattered and evaluate its performance on the primal test dataset $\mathcal{D}_{\text{test}}$. Results presented in Table 2 indicate that ART-scattered can still enhance LLMs to a certain extent, although its effectiveness has decreased somewhat compared to ART. This suggests that the choice of target attention is quite flexible. We leave the exploration of constructing better target attention for future work.

Comparison to other inference intervention methods. We compare ART with *Beam Search*, a widely used decoding strategy in LLM applications. Additionally, we evaluate a hybrid approach that combines ART with beam search, called *ART-Beam*, to investigate whether ART can further enhance LLMs. Furthermore, we compare ART with three other intervention-based methods, ITI (Li et al., 2023), ACT (Yu et al., 2024), and DoLa (Chuang et al., 2024). ITI uses in-domain training data to train probes for every attention head in the language model to identify the *Truthfulness Head*, which are then edited during inference. ACT leverages training data to identify attention heads suitable for attention calibration and subsequently edits these attention heads on the fly during inference. DoLa enhances the model truthfulness by contrasting the differences in logits obtained from final layers versus premature layers. Table 3 shows that ART can be combined with beam search for further improvement and achieves comparable or even superior performance to ITI, ACT, and DoLa on most tasks. However, unlike ITI and ACT, ART does not require task-specific downstream data for training or pilot experiments. Instead, it enhances LLMs by simply replacing attention patterns in the shallow layers, making ART more practical.

Method	Method	TruthfulQA	LogiQA	GSM8K
Llama2-7B	Vanilla	19.0	28.4	27.2
	Beam Search	21.4	29.2	29.5
	ITI	19.9	28.2	28.1
	ACT	18.7	30.3	26.7
	DoLa	21.5	27.8	28.7
	ART-greedy	21.5	30.1	28.5
	ART-beam	**22.7**	**30.5**	**30.1**
Llama3-8B	Vanilla	45.1	38.5	83.1
	Beam Search	46.0	41.0	84.3
	ITI	46.0	37.3	66.7
	ACT	45.2	38.0	72.8
	DoLa	45.3	38.8	83.2
	ART-greedy	45.8	41.4	84.0
	ART-beam	**46.4**	**42.7**	**85.8**

Table 3: Performance comparison between ART and other inference intervention methods, by evaluating *Llama2-7B-Chat* and *Llama3-8B-Instruct* on $\mathcal{D}_{\text{test}}$. **Bold** and underline denote the best and second best.

5 Conclusion

In this paper, we study attention patterns in LLMs and analyze three types of dense attention in shallow layers. Our analysis shows that local attention is important, while uniform attention is often redundant. Based on this finding, we propose **ART**, which replaces uniform attention with local attention during inference to improve attention use inside the model. Experiments on multiple open-source LLMs and datasets show that ART consistently improves downstream performance.

Limitation

It performs an empirical analysis of hallucinations and attention head patterns. However, there may be more interpretable methods that could offer deeper understanding or targeted interventions.

References

AI@Meta. 2024. Llama 3 model card.

Liwei Che, Zhiyu Xue, Yihao Quan, Benlin Liu, Zeru Shi, Michelle Hurst, Jacob Feldman, Ruixiang Tang, Ranjay Krishna, and Vladimir Pavlovic. 2026. Counting circuits: Mechanistic interpretability of visual reasoning in large vision-language models. *arXiv preprint arXiv:2603.18523*.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, volume 15139, pages 19–35. Springer.

Tianxiang Chen, Zhentao Tan, Tao Gong, Yue Wu, Qi Chu, Bin Liu, Jieping Ye, and Nenghai Yu. 2024b. Llama slayer 8b: Shallow layers hold the key to knowledge injection. In *Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5991–6002.

Tianxiang Chen, Zhentao Tan, Tao Gong, Yue Wu, Qi Chu, Bin Liu, Jieping Ye, and Nenghai Yu. 2024c. Llama slayer 8b: Shallow layers hold the key to knowledge injection. *arXiv preprint arXiv:2410.02330*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Siqi Fan, Xin Jiang, Xiang Li, Xuying Meng, Peng Han, Shuo Shang, Aixin Sun, Yequan Wang, and Zhongyuan Wang. 2024. Not all layers of llms are necessary during inference. *arXiv preprint arXiv:2403.02181*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the 41st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36 (NIPS)*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3622–3628.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32 (NIPS)*.

Yihao Quan, Zeru Shi, Jinman Zhao, and Ruixiang Tang. 2026. Reinforcing consistency in video mllms with structured rewards. *arXiv preprint arXiv:2604.01460*.

Zeru Shi, Kai Mei, Yihao Quan, Dimitris N Metaxas, and Ruixiang Tang. 2026. Improving visual reasoning with iterative evidence refinement. *arXiv preprint arXiv:2603.14117*.

Zeru Shi, Yingjia Wan, Zhenting Wang, Qifan Wang, Fan Yang, Elisa Kreiss, and Ruixiang Tang. 2025. Meaningless tokens, meaningful gains: How activation shifts enhance llm reasoning. *arXiv preprint arXiv:2510.01032*.

Qi Sun, Marc Pickett, Aakash Kumar Nain, and Llion Jones. 2024. Transformer layers as painters. *arXiv preprint arXiv:2407.09298*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4149–4158.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 5998–6008.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NIPS)*.

Jinfeng Wei and Xiaofeng Zhang. 2024a. Dopra: Decoding over-accumulation penalization and reallocation in specific weighting layer. *Proceedings of the 32nd ACM International Conference on Multimedia*.

Jinfeng Wei and Xiaofeng Zhang. 2024b. Dopra: Decoding over-accumulation penalization and reallocation in specific weighting layer. *arXiv preprint arXiv:2407.15130*.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2024a. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024b. Efficient streaming language models with attention sinks. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Fangcong Yin, Xi Ye, and Greg Durrett. 2024. Lofit: Localized fine-tuning on llm representations. *arXiv preprint arXiv:2406.01563*.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. 2024. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Xiaosong Yuan, Chen Shen, Shaotian Yan, Xiaofeng Zhang, Liang Xie, Wenxiao Wang, Renchu Guan, Ying Wang, and Jieping Ye. 2024. Instance-adaptive zero-shot chain-of-thought prompting. *arXiv preprint arXiv:2409.20441*.

Boxuan Zhang, Yi Yu, Jiaxuan Guo, and Jing Shao. 2025a. Dive into the agent matrix: A realistic evaluation of self-replication risk in llm agents. *arXiv preprint arXiv:2509.25302*.

Boxuan Zhang and Ruqi Zhang. 2025. Cot-uq: Improving response-wise uncertainty quantification in llms with chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26114–26133.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *arXiv preprint arXiv:2411.09968*.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. 2025b. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3512–3534.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2025c. From redundancy to relevance: Information flow in lvlms across reasoning tasks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2289–2299.

Xiaofeng Zhang, Fanshuo Zeng, Yihao Quan, Zheng Hui, and Jiawei Yao. 2025d. Enhancing multimodal large language models complex reason via similarity computation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10203–10211.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Jiawei Cao, Hao Cheng, and Kaijie Wu. 2026a. What drives attention sinks? a study of massive activations and rotational positional encoding in large vision–language models. *Information Processing & Management*, 63(2):104431.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, et al. 2026b. Hallucination begins where saliency drops. In *The Fourteenth International Conference on Learning Representations*.

Qiyan Zhao, Xiaofeng Zhang, Shuochen Chang, Qianyu Chen, Xiaosong Yuan, Xuhang Chen, Luoqi Liu, Jiajun Zhang, Xu-Yao Zhang, and Da-Han Wang. 2026. Context tokens are anchors: Understanding the repetition curse in dmllms from an information flow perspective. In *The Fourteenth International Conference on Learning Representations*.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2025. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

A Algorihtm Illustration

Algorithm 1 fully depicts the whole pipeline of integrating ART into the model inference. k refers to the size of the number of identified uniform attention and local attention, which is set to $k = \lfloor 0.1 * N_h \rfloor$ by default discussed in Section 4.5. L_s indicates the number of layers identified as shallow layers, which is set to $L_s = 2$ in this paper. L

represents the total number of layers the language model $\mathcal{M}$ have.

Algorithm 1 **A**ttention **R**eplacement **T**echnique

Input: question $\mathcal{Q}$, model $\mathcal{M}$, tokenizer $\mathcal{T}$, max tokens m, # uniform/local attention k, shallow layer L_s, # layer L,
Output: response $\mathcal{R}$ to question

gen_tokens $\leftarrow$ []
for each $i \in [1, m]$ **do**
  $\mathcal{X} \leftarrow$ ENCODE($\mathcal{T}, \mathcal{M}, \mathcal{Q}$, gen_tokens)
  # Forward pass: go through all L layers
  for each $l \in [1, L]$ **do**
    # Multi-Head Attention
    if $l \leq L_s$ **then**
      ART
      $\mathcal{X} \leftarrow \mathcal{X} + \text{ART-MHA}^l_{\mathcal{M},k}(\text{LN}(\mathcal{X}))$
    else
      Vanilla MHA
      $\mathcal{X} \leftarrow \mathcal{X} + \text{MHA}^l_{\mathcal{M}}(\text{LN}(\mathcal{X}))$
    end if
    # Feed-Forward Network
    $\mathcal{X} \leftarrow \mathcal{X} + \text{FFN}^l_{\mathcal{M}}(\text{LN}(\mathcal{X}))$
  end for
  $\mathcal{P} \leftarrow$ LM_HEAD($\mathcal{M}, \mathcal{X}$)
  $t_i \leftarrow$ DECODE($\mathcal{T}, \mathcal{P}$)
  gen_tokens.append(t_i)
  if stopping criteria satisfied **then**
    break
  end if
end for
return TOKENS_TO_TEXT($\mathcal{T}$, gen_tokens)

As descried in Algorithm 1, ART is only applied to the multi-head attention module of $\mathcal{M}$ and activated when the current layer is shallow. When ART is not activated, $\mathcal{M}$ performs vanilla autoregressive decoding.