Mechanistically Understanding and Mitigating the Anchoring Effect in LLMs via Attention Head Intervention

Yoshiki Takenami **Yin Jou Huang** **Yugo Murawaki** **Koh Takeuchi** **Chenhui Chu**
Kyoto University
tnami@nlp.ist.i.kyoto-u.ac.jp {huang,murawaki,takeuchi,chu}@i.kyoto-u.ac.jp

Abstract

Similarly to humans, LLMs are susceptible to the anchoring effect in price negotiations, which can lead to unfavorable or unintended outcomes. To ensure LLM safety in dialogue tasks such as negotiation, a mechanistic understanding and mitigation methods are essential. We propose four variants of a method to identify **Anchoring Heads**, which are the attention heads responsible for the anchoring effect, while preserving those essential for general tasks. Experimental results show that suppressing these heads mitigates the anchoring effect by over 16%, achieving complete mitigation in certain models without compromising general performance. Furthermore, the observation that suppressing random heads fails to mitigate the anchoring effect demonstrates that specific attention heads drive the anchoring effect by paying excessive attention to the anchor.

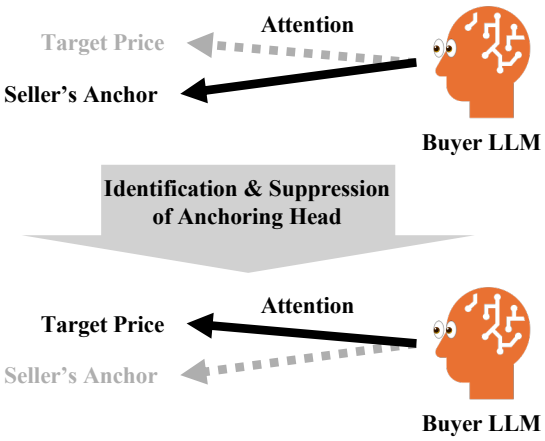


Figure 1: Identification and suppression of attention heads causing the anchoring effect in LLM-driven price negotiations. By analyzing attention weights and suppressing identified Anchoring Heads, the buyer LLM focuses more on target price instructions and mitigates the influence of the seller's anchor.

1 Introduction

Large Language Models (LLMs) possess advanced language processing capabilities; however, they have been reported to exhibit anchoring effects similar to those of humans (Sumita et al., 2025), as LLM training data reflects the biases present in human decision-making. The anchoring effect is a phenomenon where initial information serves as a reference point (an anchor), causing subsequent judgments to be unduly influenced. This effect is particularly prominent in scenarios such as price negotiations (Tversky and Kahneman, 1974).

The negotiation capabilities of LLMs have attracted attention (Liu et al., 2025; Kwon et al., 2025), and they are increasingly expected to serve as negotiation assistants. However, even when humans remain in the loop to review LLM-generated negotiation outcomes, they often fall prey to automation bias, where users over-rely on AI outputs and fail to exercise critical judgment. Therefore, to ensure the safety of LLMs in critical tasks like price negotiation, relying solely on human oversight is insufficient. It is crucial to understand the underlying internal mechanisms of the anchoring effect and develop fundamental mitigation methods.

Previous research on the anchoring effect in LLMs has explored a wide range of tasks, from numerical estimation (Suri et al., 2023) to multi-turn dialogue tasks such as negotiation (Xia et al., 2024; Takenami et al., 2025). Among these studies, Huang et al. (2025) demonstrated that specific words in a prompt, such as "higher," "lower," and the anchor value, influence the generation probability of the next token. However, their analysis remains at the behavioral level, focusing solely on input-output patterns. Although they proposed a mitigation method based on behavioral-level analysis, its effectiveness remains limited.

Beyond behavioral analysis, mechanistic interpretability has emerged as a vital framework for enhancing AI safety (Elhage et al., 2021; Nanda et al.,

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 630–644
November 6-10, 2026 ©2026 Association for Computational Linguistics

2023; Bereska and Gavves, 2024). This methodology aims to reverse-engineer the computational mechanisms and representations within neural networks into human-understandable algorithms and concepts. Regarding the anchoring effect, identifying the specific LLM components that drive these computations could facilitate the development of more robust and reliable mitigation methods.

In standard LLM architectures like the Transformer (Vaswani et al., 2017), the Feed-Forward Network serves to store factual knowledge from the training data (Geva et al., 2021; Meng et al., 2022), while the Multi-Head Attention mechanism is responsible for retrieving relevant information within context and implementing appropriate algorithms on the fly (Olsson et al., 2022; Wu et al., 2025; Todd et al., 2024). Based on this, we can hypothesize two potential mechanisms for the anchoring effect: first, that the LLM statistically imitates anchoring-related dialogue patterns present in the training data; and second, that attention heads exposed to such data attend more strongly to anchor values during inference rather than to the genuinely important information in the prompt.

Against this background, we provide a mechanistic understanding of the anchoring effect in LLM-based price negotiation simulations and propose a method for its mitigation. We hypothesize that the anchoring effect arises from specific attention heads that allocate more attention to the seller's anchor than to the target price, as illustrated by the Buyer LLM in the upper part of Figure 1. To identify such "Anchoring Heads" while excluding those broadly important for general capabilities such as conversational coherence or numerical reasoning, we propose four methods.

- Attention_Difference: Quantifies, for each head, how strongly it favors the anchor over the target price in the dialogue context.
- Attention_Difference_Shift: Examines how the attention difference shifts when a warning aginst the anchoring effect is introduced.
- Weighted_Attention_Difference: In addition to attention difference, evaluates how selectively a head concentrates its attention on particular sentences.
- Weighted_Attention_Difference_Shift: Combines Weighted_Attention_Difference and Attention_Difference_Shift.

Experimental results indicate that the most effective identification method varies across models. Attention_Difference completely eliminates the anchoring effect in Llama and achieves a 16% reduction in Phi. Attention_Difference_Shift proves effective for Qwen models, yielding mitigation rates of 33% for Qwen2.5-14B and 25% for Qwen2.5-32B. Additionally, Weighted_Attention_Difference achieves 92% mitigation in Llama, whereas Weighted_Attention_Difference_Shift does not outperform the other methods. Importantly, all methods reduce the anchoring effect without compromising performance on general tasks. Furthermore, suppressing randomly selected heads shows no such mitigation, thereby establishing a causal relationship between the identified Anchoring Heads and the anchoring effect. Our contributions are summarized as follows:

- We demonstrate that the anchoring effect is not driven by the collective cooperation of all attention heads; rather, a small fraction of the heads, which we term "anchoring heads," are responsible for this effect.
- We propose a method to mitigate the anchoring effect without compromising performance on general tasks, which achieved complete mitigation in certain models.

2 Related Work

Enhancing interpretability is essential for aligning the behavior of LLMs with human intentions and values and for ensuring safety. Interpretability can be divided into several frameworks, ranging from behavioral-level analysis that treats the model as a black box to approaches that focus on the internal structure of deep neural networks (Bereska and Gavves, 2024). Among these, Mechanistic Interpretability on LLMs has gained attention in recent years. This is a framework that identifies how individual components, such as neurons and attention heads, connect and form computational circuit to derive the final output by intervening in the internal structure of the model.

To identify the components of a computational circuit, a specific score is calculated for each individual neuron or attention head, and they are identified based on this score. Wu et al. (2025) identified retrieval heads that are largely responsible for retrieving relevant information from long context by measuring the frequency with which each attention

head copies information from the input text and pastes it into the output during text generation.

To prove the causal relationship between internal components of a model and its final output, a reliable method is to intervene in each model component and observe whether the output changes. Zhou et al. (2025) and Arai et al. (2025) manipulated each attention head and measured how much the probability distribution of the output for harmful queries changed from the original, thereby identifying attention heads that influence whether a model produces a harmful or safe response. Although this method can reliably identify the components, it is impractical for time-consuming dialogue tasks such as the price negotiation addressed in this paper.

As a lightweight alternative, there is a method of comparing internal states when paired prompts are provided. Zou et al. (2025) proposed Representation Engineering, which compares and extracts the internal states of LLMs when given opposing prompts (such as "lie" and "tell the truth") related to specific concepts (honesty, happiness, and fairness). Zhong et al. (2026) compared and extracted the internal states when an English prompt was input versus when another language was input, thereby controlling the output language.

Attention_Difference_Shift, which identifies heads that shift their attention bias depending on the presence or absence of a warning against the anchoring effect, was inspired by Zou et al. (2025); Zhong et al. (2026). While their methods analyze and extract representation, our approach identifies attention heads. Furthermore, this is the first study to address the mechanistic understanding of the anchoring effect on LLM.

3 Method

3.1 Investigation of the Anchoring Effect in LLM-driven Price Negotiation Simulation

Following the method of Takenami et al. (2025), price negotiations are conducted between two LLMs acting as a seller and a buyer. Each LLM is assigned personality traits based on the Big Five, which is controlled through adjectives in the prompts. The dialogue proceeds until an agreement, a breakdown, or the maximum turn limit is reached. To investigate the anchoring effect, we conduct simulations under the following two conditions:

- Control: A condition where no specific instructions regarding the anchoring effect are given.
- Anchoring: A condition where the seller LLM is explicitly instructed via a prompt to use the anchoring effect.

3.2 Identification of Anchoring Heads

We conduct price negotiation simulations using LLMs following the method in Section 3.1 and, as shown in Figure 2, analyze the attention when the resulting dialogue history is provided to the same LLM. The dialogue history consists of the system prompt given to the buyer LLM and the utterances of the seller and buyer LLMs.

Because it is difficult to analyze attention in long multi-turn dialogues at the token level, we analyze the attention mechanism at the sentence level. We define S_m and S_n as the sets of tokens in the m-th and n-th sentences of the dialogue history. We define the sentence-wise attention from the m-th sentence to the n-th sentence ($n \leq m-1$) at the h-th head in the l-th layer as follows:

$$A^{(l,h)}(m,n) = \frac{1}{|S_m||S_n|} \sum_{i \in S_m} \sum_{j \in S_n} \alpha_{i,j}^{(l,h)} \quad (1)$$

where $\alpha_{i,j}^{(l,h)}$ represents the attention weight from the i-th token in the m-th sentence to the j-th token in the n-th sentence.

To evaluate how sharply an attention head focuses on specific sentences, following Bi et al. (2025), we calculate the sentence-wise attention entropy $H^{(l,h)}(m)$ when processing the m-th sentence:

$$H^{(l,h)}(m) = -\sum_{n=1}^{m-1} \frac{A^{(l,h)}(m,n)}{A^{(l,h)}(m)} \times \log_2 \frac{A^{(l,h)}(m,n)}{A^{(l,h)}(m)} \quad (2)$$

Here, $A^{(l,h)}(m)$ is the sum of inter-sentence attention when processing the m-th sentence.

$H^{(l,h)}(m)$ has no fixed upper bound, as its maximum value depends on m. Therefore, we normalized it by its theoretical maximum. Furthermore, since our objective is to detect attention heads that respond sensitively to the anchor sentence or the target price instruction, we defined concentration $C^{(l,h)}(m)$ by subtracting the normalized entropy from 1. This ensures that the concentration value is positively correlated with the sensitivity of the head.

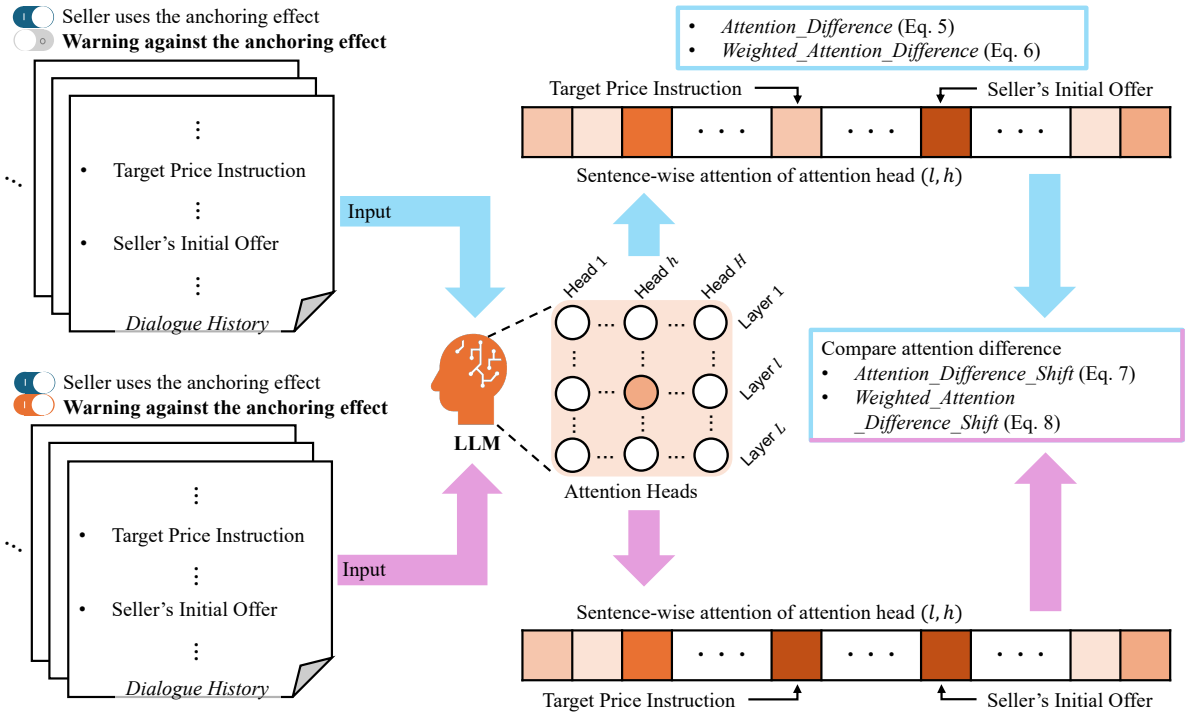


Figure 2: Method for identifying Anchoring Heads. We analyze the attention weights when an LLM processes the dialogue history of price negotiations conducted by LLMs. The proposed method has four variations: methods that detect heads paying stronger attention or concentration-weighted attention to the seller's initial offer than to the target price instruction (Attention_Difference, Weighted_Attention_Difference), and methods that detect heads whose behavior changes depending on the presence or absence of the *warning sentence* (Attention_Difference_Shift, Weighted_Attention_Difference_Shift).

$$C^{(l,h)}(m) = 1 - \frac{H^{(l,h)}(m)}{\log_2(m-1) + \epsilon} \quad (3)$$

ϵ is a small constant to prevent zero division when $m \leq 2$.

To amplify the weights of heads exhibiting high concentration, we define the concentration-weighted sentence-wise attention $\tilde{A}^{(l,h)}(m,n)$:

$$\tilde{A}^{(l,h)}(m,n) = A^{(l,h)}(m,n)(1 + C^{(l,h)}(m)) \quad (4)$$

To quantify the bias toward the anchor value, we define n_1 as the anchor sentence and n_2 as the target price sentence. We compute the attention difference $D^{(l,h)}(n_1,n_2)$ and the concentration-weighted attention difference $\tilde{D}^{(l,h)}(n_1,n_2)$ across the entire dialogue history of M sentences.

$$D^{(l,h)}(n_1,n_2) = \frac{1}{M - M_{\text{start}}} \sum_{m=M_{\text{start}}}^{M} \left(A^{(l,h)}(m,n_1) - A^{(l,h)}(m,n_2) \right) \quad (5)$$

$$\tilde{D}^{(l,h)}(n_1,n_2) = \frac{1}{M - M_{\text{start}}} \sum_{m=M_{\text{start}}}^{M} \left(\tilde{A}^{(l,h)}(m,n_1) - \tilde{A}^{(l,h)}(m,n_2) \right) \quad (6)$$

where $M_{\text{start}} = \max(n_1,n_2) + 1$. A larger attention difference indicates that the head attends more strongly to the anchor sentence than to the target price sentence.

To identify heads that selectively react to the anchoring effect, we compute the shift in attention behavior across K dialogue histories when the *warning sentence* is provided (informed) versus when it is not (uninformed). The *warning sentence* is a prompt that alerts the buyer LLM that the seller LLM will employ the anchoring effect. The specific prompt and its ablation study are provided in Appendix A.

$$\Delta D^{(l,h)}(n_1, n_2) = \frac{1}{K} \sum_{k=1}^{K} \left(D^{(l,h)}_{\text{uninformed},k}(n_1, n_2) - D^{(l,h)}_{\text{informed},k}(n_1, n_2) \right) \tag{7}$$

$$\Delta \tilde{D}^{(l,h)}(n_1, n_2) = \frac{1}{K} \sum_{k=1}^{K} \left(\tilde{D}^{(l,h)}_{\text{uninformed},k}(n_1, n_2) - \tilde{D}^{(l,h)}_{\text{informed},k}(n_1, n_2) \right) \tag{8}$$

A higher score indicates that the head exhibits a shift in its attention distribution in response to the *warning sentence*, decreasing its focus on the anchor and increasing its focus on the target price. However, it should be noted that while these methods can selectively identify heads reacting to the anchoring effect, they may also capture heads that actively work to resist it.

To identify Anchoring Heads, we propose and compare four methods, selecting the top k heads by sorting them in descending order of their respective scores:

- Attention_Difference: $D^{(l,h)}(n_1, n_2)$
- Attention_Difference_Shift: $\Delta D^{(l,h)}(n_1, n_2)$
- Weighted_Attention_Difference: $\tilde{D}^{(l,h)}(n_1, n_2)$
- Weighted_Attention_Difference_Shift: $\Delta \tilde{D}^{(l,h)}(n_1, n_2)$

3.3 Suppression of Attention Heads

To verify whether the Anchoring Heads identified by the proposed method play a causal role in the occurrence of the anchoring effect, we propose to suppress the function of these heads. In the Multi-Head Attention mechanism of the l-th layer with N_H query heads, the output $Y^{(l)}$ for the input X can be formulated as a weighted sum of the output of each head, using the output $H^{(l,h)}$ of each head h and the weight matrix $W_O^{(l,h)}$, as shown in the following equation.

$$Y^{(l)} = \sum_{h=1}^{N_H} H^{(l,h)} W_O^{(l,h)} \tag{9}$$

For each head $(l, h) \in \mathcal{H}$ included in the set of Anchoring Heads $\mathcal{H}_{\text{anc}}$, we suppress its contribution by multiplying the corresponding output weight $W_O^{(l,h)}$ by a small coefficient α. The weight after the intervention $\hat{W}_O^{(l,h)}$ is given by the following equation.

$$\hat{W}_O^{(l,h)} = \begin{cases} \alpha W_O^{(l,h)} & \text{if } (l,h) \in \mathcal{H}_{\text{anc}} \\ W_O^{(l,h)} & \text{otherwise} \end{cases} \tag{10}$$

4 Experiment

4.1 Setting

We employed several open-weight models capable of successful price negotiation, including Qwen2.5-32B,[1] Qwen2.5-14B,[2] Llama-3.1-8B,[3] and phi-4.[4] We conducted the price negotiation simulations by fixing the seller LLM to Qwen2.5-32B, the largest model, with the expectation that it would perform stable price negotiations regardless of the buyer. To observe the impact of head suppression, we replaced the buyer LLM with a model in which specific heads were suppressed. To identify the anchor sentence in the dialogue history and evaluate the outputs on MT-Bench, we used GPT-4.1.[5]

Following Huang et al. (2025); Takenami et al. (2025), we extracted 161 product information entries (negotiation scenarios involving various products and target prices) from the CraigslistBargains dataset (He et al., 2018), assigning two pairs of personality traits (one for buyer and one for seller) for each scenario. This resulted in a total of 322 negotiation scenarios. We utilized 128 of these scenarios to identify the Anchoring Heads and another 64 for hyperparameter tuning of parameters k and α. The evaluation results for the remaining 130 scenarios are presented in Section 4.4.

4.2 Evaluation Metric

We calculate the achievement degree of the buyer's target price:

$$\text{Achievement}(p) = \frac{\overline{p_b} - p}{\overline{p_b} - \underline{p_b}} \tag{11}$$

Here, $\overline{p}_b$ and $\underline{p}_b$ denote the buyer's maximum acceptable price and target price, respectively. The

[1] Qwen/Qwen2.5-32B-Instruct
[2] Qwen/Qwen2.5-14B-Instruct
[3] meta-llama/Llama-3.1-8B-Instruct
[4] microsoft/phi-4
[5] gpt-4.1-2025-04-14

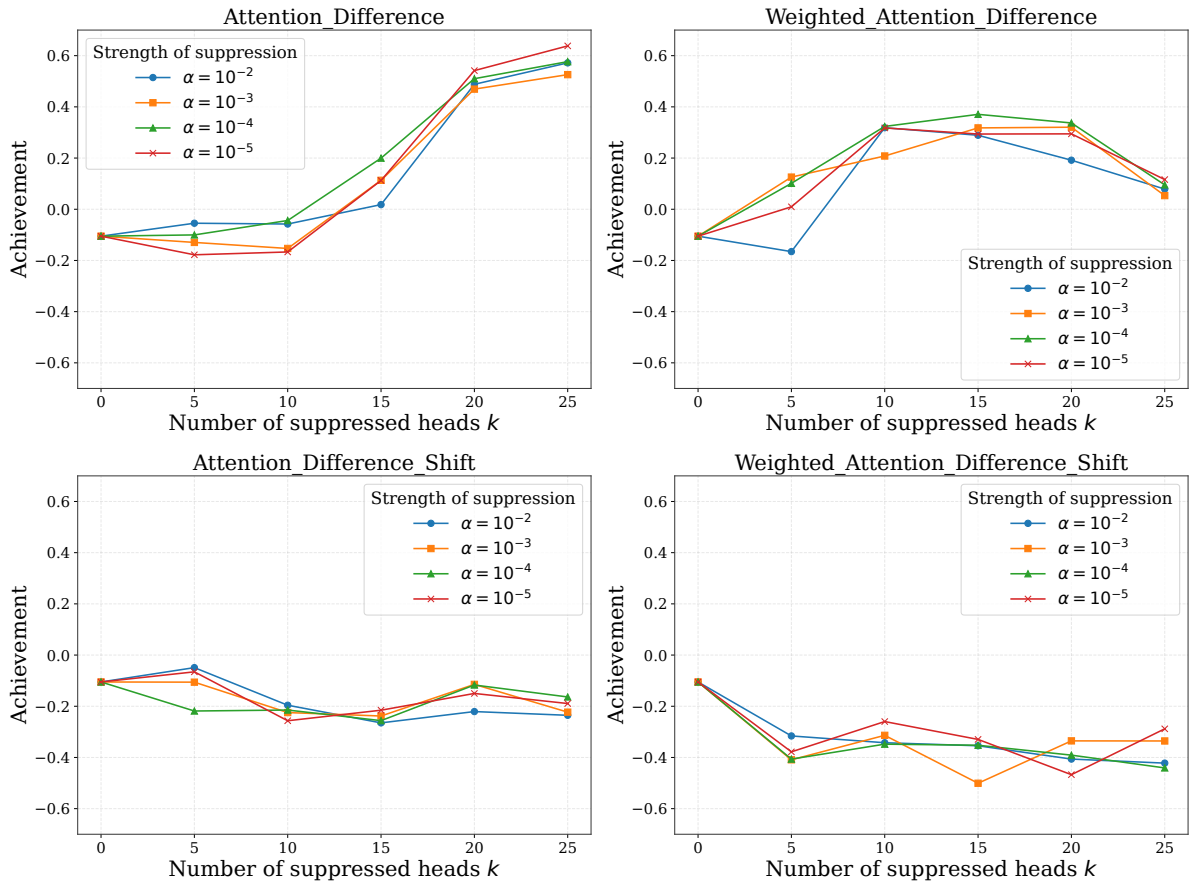


Figure 3: Grid search of hyperparameters for Llama-3.1-8B. The vertical axis represents the achievement degree of the buyer LLM's target price, where a larger value indicates that the anchoring effect was mitigated. For Llama-3.1-8B, the Attention_Difference method was the most effective, but as shown in the Appendix B, the optimal method depends on the model.

maximum acceptable price $\overline{p}_b$ is heuristically calculated to divide the seller's target price and the buyer's target price at a ratio of $3:7$, and is not provided to the LLMs during the simulation. A lower achievement indicates that the buyer suffered a greater loss due to the anchoring effect of the seller.

Furthermore, to investigate the impact of head suppression on other tasks, we utilized MMLU (Hendrycks et al., 2021) to measure knowledge and problem-solving abilities, and MT-Bench (Zheng et al., 2023) to evaluate dialogue and instruction-following capabilities.

4.3 Hyperparameter Tuning

For each model and method, we conducted simulations under the Anchoring condition. In this section, we use Llama-3.1-8B as an example. Please refer to the appendix B for the other models. As shown in Figure 3, we performed a grid search with $k = \{5, 10, 15, 20, 25\}$ and $\alpha = \{0.01, 0.001, 0.0001, 0.00001\}$. The vertical axis represents Achievement, and a larger value compared to the LLM without head suppression (the leftmost point in each subfigure) indicates that the anchoring effect was mitigated. We selected the hyperparameters that most effectively mitigated the anchoring effect, as shown in Table 1.

Method	k	α
Attention_Difference	25	0.00001
Attention_Difference_Shift	5	0.01
Weighted_Attention_Difference	15	0.0001
Weighted_Attention_Difference_Shift	10	0.00001

Table 1: Tuned hyperparameters for Llama-3.1-8B. See Appendix B for other models.

4.4 Result

We conducted price negotiation simulations by suppressing the Anchoring Heads identified by each of the four proposed methods with the tuned hyper-

	achievement ↑		MMLU	MT-Bench
	Control	Anchoring		
Qwen2.5-14B-Instruct	$0.23_{\pm 0.47}$	$-0.20_{\pm 0.58}$	0.79	8.27
+ Attention_Difference	N/A	$-0.12_{\pm 0.61}$	0.78	8.18
+ Attention_Difference_Shift	N/A	$\mathbf{-0.06_{\pm 0.56}}$	0.79	8.26
+ Weighted_Attention_Difference	N/A	$\underline{-0.08_{\pm 0.61}}$	0.79	8.26
+ Weighted_Attention_Difference_Shift	N/A	$-0.23_{\pm 0.57}$	0.78	8.50
Qwen2.5-32B-Instruct	$0.11_{\pm 0.49}$	$-0.44_{\pm 0.54}$	0.83	8.54
+ Attention_Difference	N/A	$-0.40_{\pm 0.54}$	0.83	8.18
+ Attention_Difference_Shift	N/A	$\mathbf{-0.30_{\pm 0.60}}$	0.82	8.30
+ Weighted_Attention_Difference	N/A	$-0.51_{\pm 0.63}$	0.83	8.50
+ Weighted_Attention_Difference_Shift	N/A	$-0.37_{\pm 0.51}$	0.83	8.49
Llama-3.1-8B-Instruct	$0.36_{\pm 0.49}$	$-0.15_{\pm 0.73}$	0.69	6.72
+ Attention_Difference	N/A	$\mathbf{0.44_{\pm 0.69}}$	0.63	6.49
+ Attention_Difference_Shift	N/A	$-0.10_{\pm 0.65}$	0.67	6.91
+ Weighted_Attention_Difference	N/A	$\underline{0.32_{\pm 0.62}}$	0.65	6.68
+ Weighted_Attention_Difference_Shift	N/A	$-0.32_{\pm 0.66}$	0.57	6.75
phi-4	$0.24_{\pm 0.49}$	$-0.31_{\pm 0.63}$	0.81	7.41
+ Attention_Difference	N/A	$-0.22_{\pm 0.65}$	0.79	7.20
+ Attention_Difference_Shift	N/A	$-0.36_{\pm 0.55}$	0.80	7.55
+ Weighted_Attention_Difference	N/A	$-0.41_{\pm 0.69}$	0.79	7.29
+ Weighted_Attention_Difference_Shift	N/A	$-0.30_{\pm 0.60}$	0.81	7.48

Table 2: The means and standard deviations of achievement of target prices by buyer LLMs under two conditions: Control (normal negotiation), Anchoring (seller LLM uses the anchoring effect). Achievement measures the extent to which the final agreed price achieved the target price. Bold and underlined values indicate the highest and second-highest statistically significant results within each model, respectively. Among the four proposed methods for attention head suppression, Attention_Difference_Shift was most effective for the Qwen family, while Attention_Difference was most effective for Llama-3.1-8B and phi-4 in mitigating the anchoring effect. As shown by the results on MMLU and MT-Bench, performance on general tasks was not affected.

parameters. Table 2 shows the means and standard deviations of achievement scores under the Control and Anchoring, along with the evaluation results on MMLU and MT-Bench. For details of the statistical analysis and the analysis of the layer distribution of the Anchoring Heads, please refer to Appendix C and Appendix D, respectively.

First, focusing on models without suppression, achievement in the Anchoring condition is lower than in the Control condition. This indicates that the buyer LLM was influenced by the seller LLM's anchor, resulting in an agreement at a higher price. Next, we report the percentage by which this anchoring effect was mitigated through the suppression of Anchoring Heads.

The optimal method varies by model. For the Qwen family, Attention_Difference_Shift was most effective, reducing the anchoring effect by 33% for Qwen2.5-14B and 25% for Qwen2.5-32B. For Llama-3.1-8B, Attention_Difference and Weighted_Attention_Difference were effective, achieving reduction rates of over 100% and 92%, respectively. For phi-4, Attention_Difference was effective, mitigating the anchoring effect by 16%. Furthermore, little to no impact was observed on either MMLU or MT-Bench, suggesting that the heads identified by our proposed method do not include those essential for general tasks.

For Attention_Difference in Llama, which achieved complete mitigation of the anchoring effect, we analyzed 15 out of 130 dialogue histories. In 7 of these 15 cases, it was observed that the model with suppression reached lower agreement prices by repeatedly producing utterances that adhered to the target price.

To detect attention heads that focus sharply on the anchor or target price sentences, we introduced the concept of concentration. However, experimental results showed Weighted_Attention_Difference and Weighted_Attention_Difference_Shift were only effective for Qwen2.5-14B and Llama. By definition, concentration takes a large value even when a head focuses sharply on sentences other than the anchor or target price. Given that attention mechanisms possess a property known as "attention sinks," where high attention is directed toward the beginning of a sequence or special tokens (Xiao et al., 2024), it is possible that the concentration-

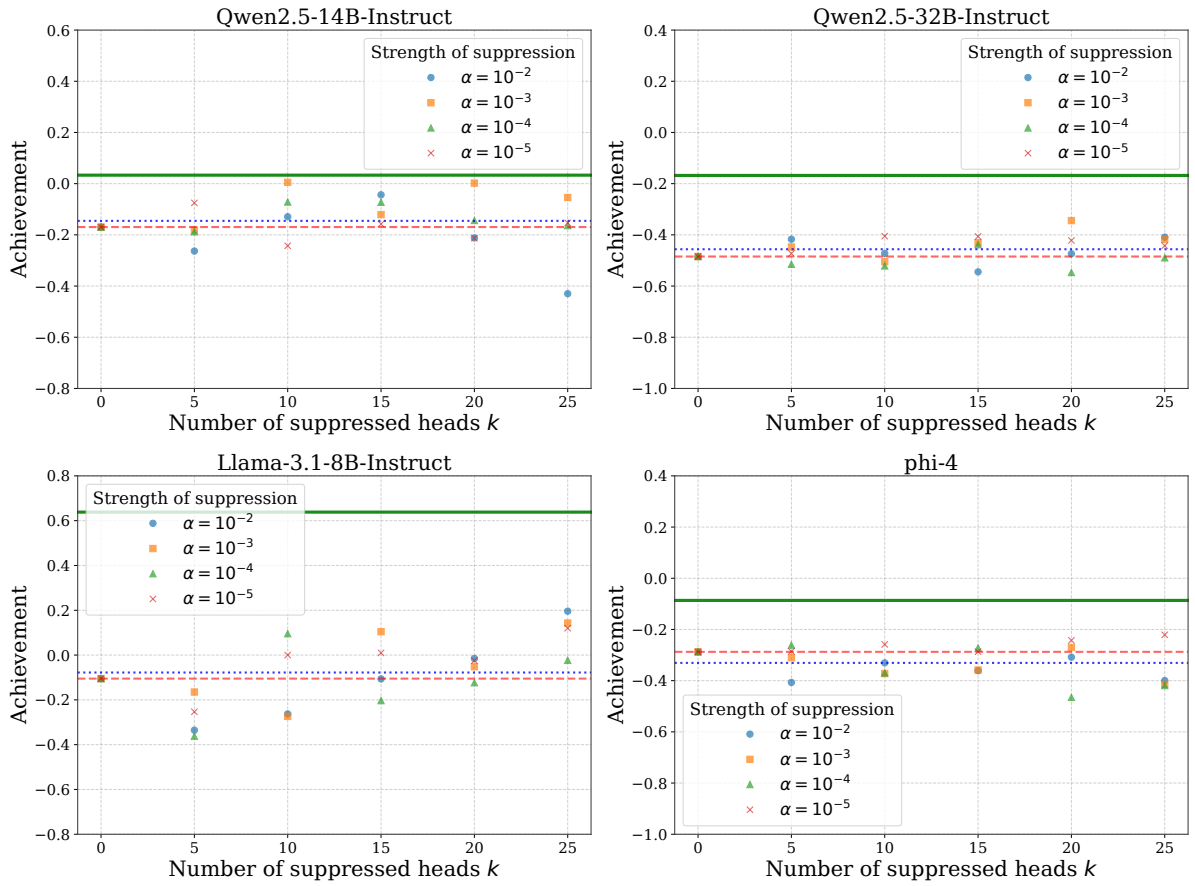


Figure 4: Target price achievement of models with randomly suppressed heads, where heads were stochastically selected for each combination of k and α. Notably, none of the random trials reached the performance of the proposed method (—). Even the highest-performing random trials failed to match the achievement of the best proposed method for each LLM. The overall average of these random trials (····) remained nearly identical to the baseline (- -).

based methods detected heads unrelated to the anchoring effect.

5 Causal Validation via Random Head Suppression

For each LLM, we performed 20 trials of price negotiation simulations under the Anchoring condition by randomly selecting heads for each combination of k and α, where $k \in \{5, 10, 15, 20, 25\}$ and $\alpha \in \{0.01, 0.001, 0.0001, 0.00001\}$. The results are shown in Figure 4. None of the LLMs with randomly suppressed heads reached the maximum performance achieved by the LLMs with suppressed Anchoring Heads, with the average value converging to nearly the same level as that of the unsuppressed LLMs. Furthermore, the large variance in achievement observed in Qwen2.5-14B and Llama-3.1-8B is likely because smaller models are more sensitive even to the suppression of a small number of heads.

Because the suppression of Anchoring Heads mitigates the anchoring effect, whereas random head suppression does not yield reliable results, a causal link between Anchoring Heads and the anchoring effect is established. Specific attention heads that prioritize the seller's anchor over the buyer's target price were mechanistically identified as components of the circuit underlying the anchoring effect.

6 Conclusion

In this study, we propose four methods to identify specific attention heads responsible for the anchoring effect in LLMs during price negotiations, providing a mechanistic understanding of this bias. Our findings lay the foundation for enhancing LLM safety in negotiation tasks. Models with suppressed Anchoring Heads achieved an over 16% reduction in the anchoring effect, with complete mitigation in certain models, without compromising perfor-

mance on general tasks. The fact that random head suppression failed to mitigate the anchoring effect demonstrates the successful identification of the specific circuit driving this bias in LLMs.

One promising direction for future research is to investigate the influence of training on human negotiation corpora or human preference data. Specifically, both how the Anchoring Heads emerge and are distributed during training, as well as why certain methods succeed while others do not, are heavily dependent on the training data and training recipe.

Limitations

- We did not analyze attention weights dynamically during the price negotiation itself. This is because we utilized vLLM (Kwon et al., 2023) to improve the throughput of the LLMs, and vLLM does not support the output of attention weights. Since multi-turn simulations like price negotiation are highly time-consuming, the use of vLLM was indispensable. Therefore, we leave the real-time analysis of attention to future work focusing on anchoring effects in single-turn dialogues. Furthermore, the dialogue history was utilized exclusively during the score calculation of our proposed methods. For all other phases, including hyperparameter tuning and the main results presented in Table 2, the experiments were conducted via simulations that closely mirror real-world applications.

- We focuse exclusively on the attention mechanism and does not address Feed-Forward Networks or other components. While completely identifying the circuit of the anchoring effect in LLMs might potentially allow for its total mitigation, the buyer LLM would, in that case, be ignoring the seller LLM's utterances. An LLM that ignores its partner's utterances or instructions is no longer useful; thus, it is essential to balance the mitigation of the anchoring effect with the preservation of instruction-following capabilities.

- As described in Section 4.1, our experimental design focused on analyzing the internal attention mechanisms of the buyer LLM. To accurately isolate the impact of suppressing specific attention heads in the buyer, we needed a highly controlled environment. By fixing the seller LLM to Qwen2.5-32B, which is the largest model we employed, we aimed to ensure stable price negotiations regardless of the buyer's behavior. This design minimized the confounding variable of seller output variance, allowing us to purely measure the efficacy of our attention head intervention across multiple different buyer models, including Qwen2.5-14B, Qwen2.5-32B, Llama-3.1-8B, and phi-4. Our current findings and mitigation success rates are based on interactions with a single, highly capable seller model (Qwen2.5-32B).

Ethical Considerations

We maintain a negative stance toward obtaining unfair profits through the use of LLMs in price negotiation. The objective of this study is to clarify the mechanisms underlying the anchoring effect in LLMs, and we hope that the findings of this research will be received responsibly and ethically.

Based on our original text and drafts, Generative AI was utilized for portions of the English translation and coding. We have reviewed all AI-generated outputs and take full responsibility for the content.

Acknowledgements

This work was supported by JST CREST (JPMJCR20D2), JST FOREST (JPMJFR232S), and JSPS KAKENHI (JP23K28144).

References

Masaki Arai, Toshiki Shibahara, Daiki Chiba, Mitsuaki Akiyama, and Masato Uchida. 2025. Jailbreak Defense in LLM via Attention Head Analysis and Selective Intervention. In *The 17th Asian Conference on Machine Learning (Conference Track)*.

Leonard Bereska and Stratis Gavves. 2024. Mechanistic Interpretability for AI Safety - A Review. *Transactions on Machine Learning Research*. Survey Certification, Expert Certification.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. 2025. Unveiling Visual Perception in Language Models: An Attention Head Analysis Approach. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4135–4144. IEEE.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others.

2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.

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.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling Strategy and Generation in Negotiation Dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Yiming Huang, Biquan Bie, Zuqiu Na, Weilin Ruan, Songxin Lei, Yutao Yue, and Xinlei He. 2025. An Empirical Study of the Anchoring Effect in LLMs: Existence, Mechanism, and Potential Mitigations. *Preprint*, arXiv:2505.15392.

Deuksin Kwon, Jiwon Hae, Emma Clift, Daniel Shamsoddini, Jonathan Gratch, and Gale Lucas. 2025. ASTRA: A Negotiation Agent with Adaptive and Strategic Reasoning via Tool-integrated Action for Dynamic Offer Optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16217–16238, Suzhou, China. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. *Preprint*, arXiv:2309.06180.

Yutong Liu, Lida Shi, Rui Song, and Hao Xu. 2025. A Dual-Mind Framework for Strategic and Expressive Negotiation Agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23840–23860, Vienna, Austria. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

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

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context Learning and Induction Heads. *Preprint*, arXiv:2209.11895.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. 2025. Cognitive Biases in Large Language Models: A Survey and Mitigation Experiments. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, SAC '25, page 1009–1011, New York, NY, USA. Association for Computing Machinery.

Gaurav Suri, Lily R. Slater, Ali Ziaee, and Morgan Nguyen. 2023. Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5. *Preprint*, arXiv:2305.04400.

Yoshiki Takenami, Yin Jou Huang, Yugo Murawaki, and Chenhui Chu. 2025. How Does Cognitive Bias Affect Large Language Models? A Case Study on the Anchoring Effect in Price Negotiation Simulations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4481–4498, Suzhou, China. Association for Computational Linguistics.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. Function Vectors in Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2025. Retrieval Head Mechanistically Explains Long-Context Factuality. In *The Thirteenth International Conference on Learning Representations*.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. 2024. Measuring Bargaining Abilities of LLMs: A Benchmark and A Buyer-Enhancement Method. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3579–3602, Bangkok, Thailand. Association for Computational Linguistics.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient Streaming Language Models with Attention Sinks. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Yugo Murawaki, Chenhui Chu, and Sadao Kurohashi. 2026. Language Lives in Sparse Dimensions: Toward Interpretable and Efficient Multilingual Control for Large Language Models. *Preprint*, arXiv:2510.07213.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. 2025. On the Role of Attention Heads in Large Language Model Safety. In *The Thirteenth International Conference on Learning Representations*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation Engineering: A Top-Down Approach to AI Transparency. *Preprint*, arXiv:2310.01405.

A Warning Sentence and its Ablation Study

Figure 5 presents the warning sentence against the anchoring effect used for Attention_Difference_Shift and Weighted_Attention_Difference_Shift. It consists of a notification that the seller will employ the anchoring effect, a definition of the anchoring effect, and countermeasures.

However, the warning sentence shown in Figure 5 is quite long and directly instructs the model to engage in multiple distinct behaviors. Consequently, there is a possibility that the heads identified using this warning sentence are not specific to the anchoring effect itself, but rather are merely responding to the instructions within a lengthy prompt. To investigate this possibility, we simplified the warning sentence by removing the countermeasures, and then identified and suppressed the Anchoring Heads in Qwen2.5-32B. The results of the price negotiation simulations are shown in Figure 6. When using the simplified warning sentence (the dark lines), the achievement is lower than when using the original warning sentence (the light lines). This demonstrates that our method using the original warning sentence successfully identifies the heads responsible for the anchoring effect, rather than those merely responding to a lengthy prompt that includes countermeasures.

> Be aware that the seller will be using the anchoring effect, a cognitive bias where the initial offer greatly influences the outcome of the negotiation. In this case, the seller will start with an initial offer that is higher than their actual target price, aiming to set an "anchor" that will shape your expectations and potentially increase the final agreed price.
>
> notification, definition
>
> To effectively negotiate under the influence of the anchoring effect, you can take the following strategies.
> 1. Stay Focused on Your Target Price: Remember your original budget or target price, and use it as a reference point instead of being swayed by the seller's initial high offer.
> 2. Set a Counter-Anchor if Needed: If the seller's initial offer is significantly higher than your budget, consider responding with a counter-offer that's closer to your ideal price. This can help shift the final agreed price closer to your preferred range.
> 3. Ask for Justification of the High Price: Politely inquire about the specifics that justify the seller's high initial offer. This can provide context for the price and might allow you to negotiate on specific elements, such as additional benefits or discounts.
>
> countermeasures

Figure 5: The warning sentence used for the proposed methods comprises a notification, a definition, and countermeasures. In the ablation study presented in Appendix A, we utilized a simplified warning sentence with the countermeasures removed.

B Hyperparameter Tuning

Figures 7 to 9 show the experimental results for each combination of hyperparameters k and α under the Anchoring condition, and Tables 3 to 5 present the tuning results for Qwen2.5-14B, Qwen2.5-32B, and Phi-4, respectively.

Method	k	α
Attention_Difference	25	0.00001
Attention_Difference_Shift	25	0.0001
Weighted_Attention_Difference	15	0.01
Weighted_Attention_Difference_Shift	25	0.01

Table 3: Tuned hyperparameters for Qwen2.5-14B.

C Statistical Analysis

In this section, we explain the standard deviations in Table 2 and the conditions for the statistical tests. The standard deviations in Table 2 are large enough to be comparable to the difference in mean values between conditions (Control and Anchoring). This

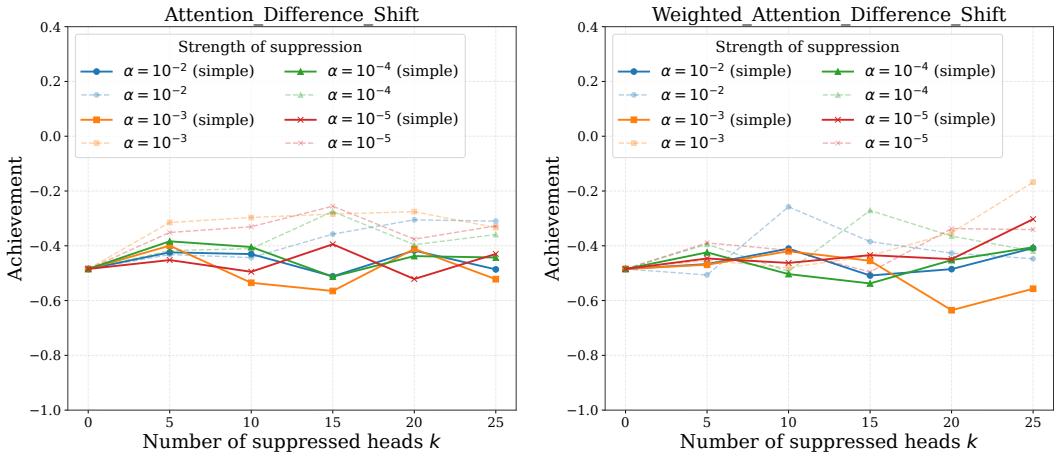


Figure 6: The results of the ablation study on the warning sentence demonstrate that the original warning sentence (the light lines) yields a higher achievement than the simplified warning sentence (the dark lines). This confirms that our method using the original warning sentence possesses a superior capability in identifying the Anchoring Heads.

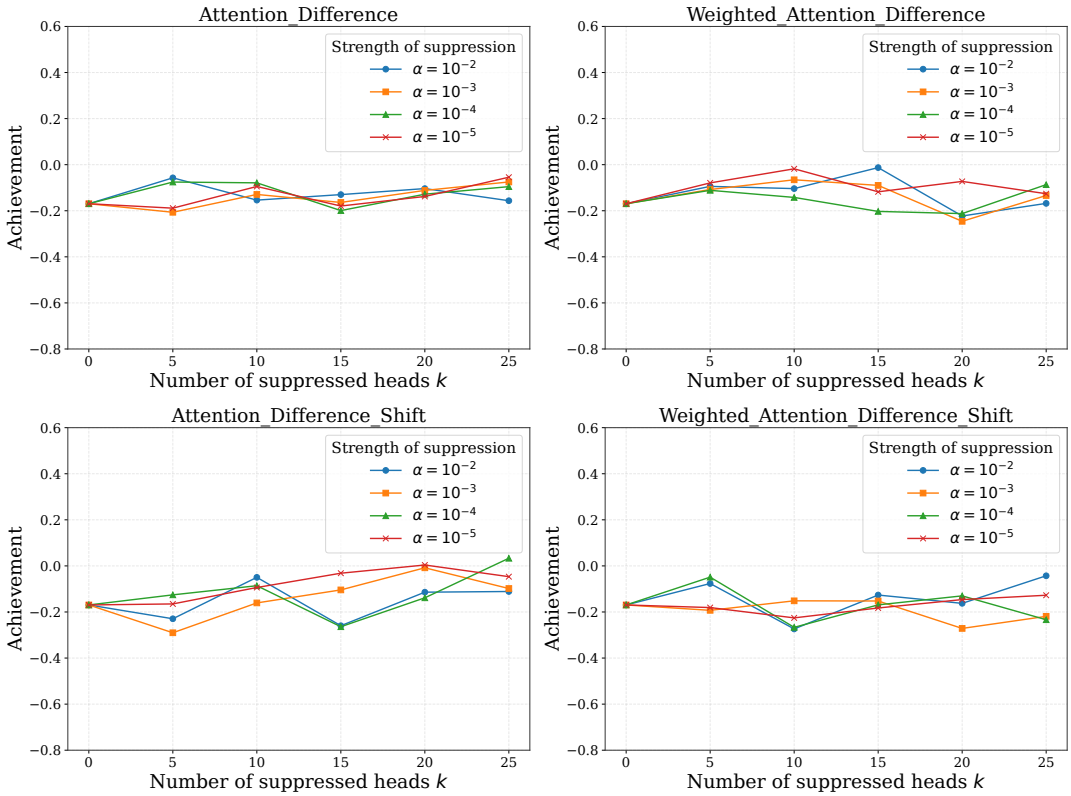


Figure 7: Hyperparameter tuning on Qwen2.5-14B

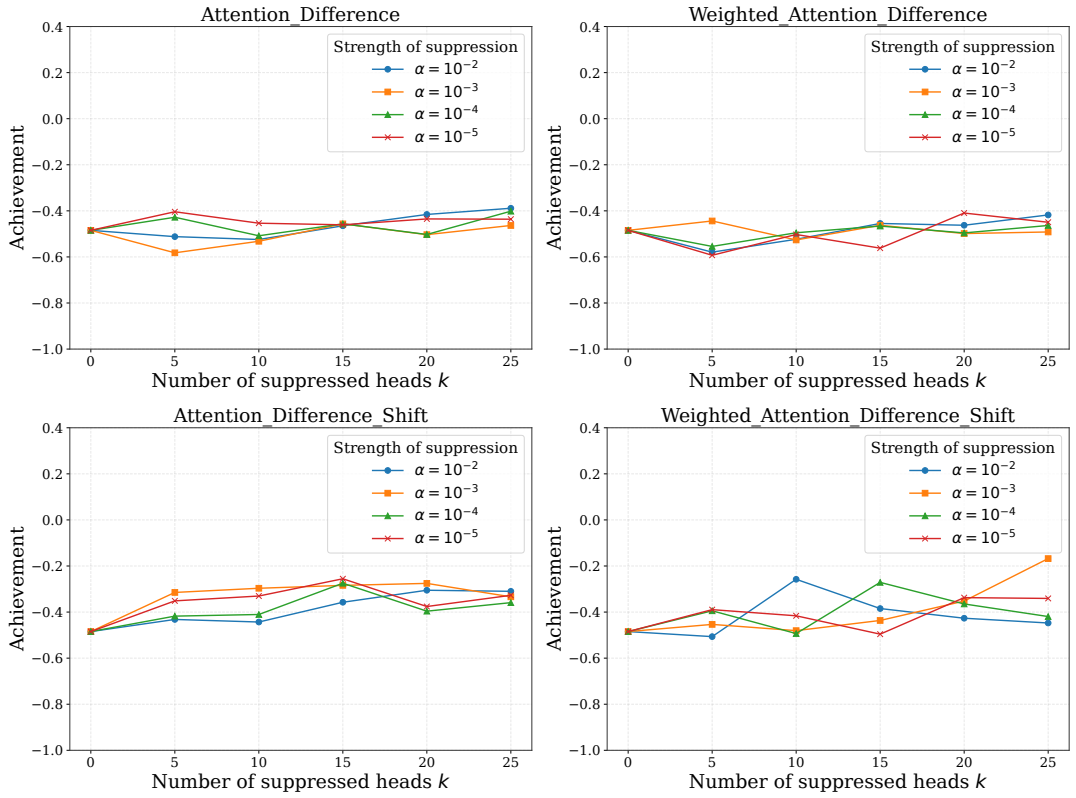


Figure 8: Hyperparameter tuning on Qwen2.5-32B

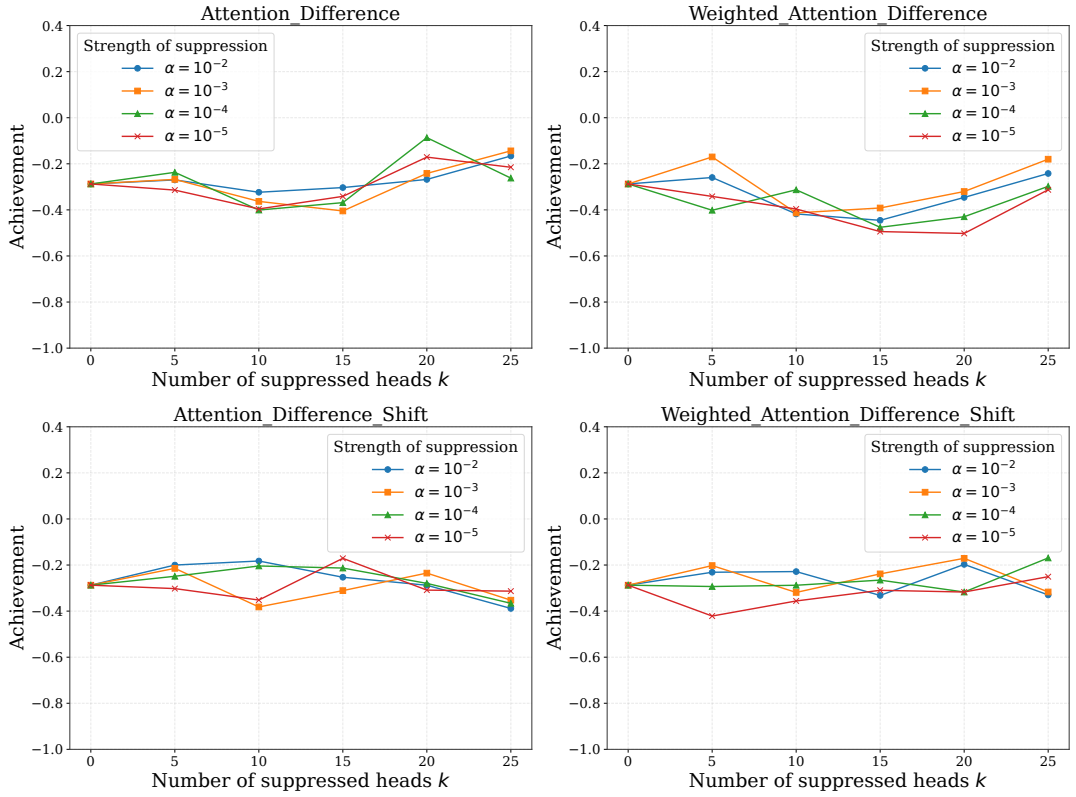


Figure 9: Hyperparameter tuning on phi-4

Method	k	α
Attention_Difference	25	0.01
Attention_Difference_Shift	15	0.00001
Weighted_Attention_Difference	20	0.00001
Weighted_Attention_Difference_Shift	25	0.001

Table 4: Tuned hyperparameters for Qwen2.5-32B

Method	k	α
Attention_Difference	20	0.0001
Attention_Difference_Shift	15	0.00001
Weighted_Attention_Difference	5	0.001
Weighted_Attention_Difference_Shift	20	0.0001

Table 5: Tuned hyperparameters for phi-4

is primarily due to the inclusion of negotiation simulations for low-priced products. For inexpensive items such as a stool, where the buyer's target price and maximum budget are set at 13 and 15, respectively, even a slight difference in the final agreed price leads to significant fluctuations in the achievement value.

For the statistical significance tests, we conducted a one-tailed paired t-test with the null hypothesis being that "the mean achievement of the LLM with heads suppressed (using each proposed method) is equal to the mean achievement of the vanilla LLM under the Anchoring condition." The resulting p-values are summarized in Table 6.

D Distribution of Identified Heads

We analyzed the distribution of the top 25 heads identified by the four proposed methods (See Figures 10 to 13 in Appendix D for the distribution of Anchoring Heads identified by each proposed method). In Qwen2.5-14B and Qwen2.5-32B, Attention_Difference_Shift identified heads fairly evenly across most layers with the exception of the 20–40% range. In contrast, for Llama-3.1-8B, Attention_Difference primarily identified heads of the first 20% layers, whereas for Phi-4, they were predominantly concentrated within the 20–60% range. These model-dependent variations in head distribution suggest that a more detailed analysis, accounting for architectural designs and training datasets, is required.

Figures 10 to 13 show the distribution of Anchoring Heads identified by each proposed method for Qwen2.5-14B, Qwen2.5-32B, Llama-3.1-8B, and Phi-4, respectively.

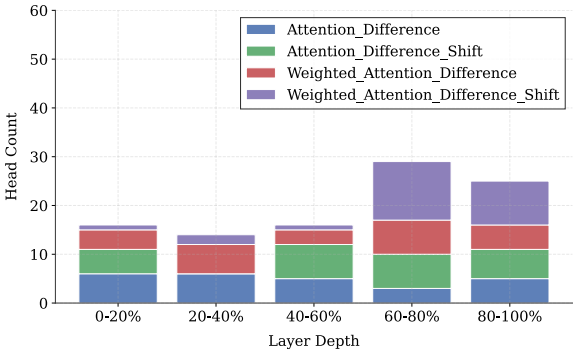


Figure 10: Distribution of the heads identified by the four proposed methods on Qwen2.5-14B

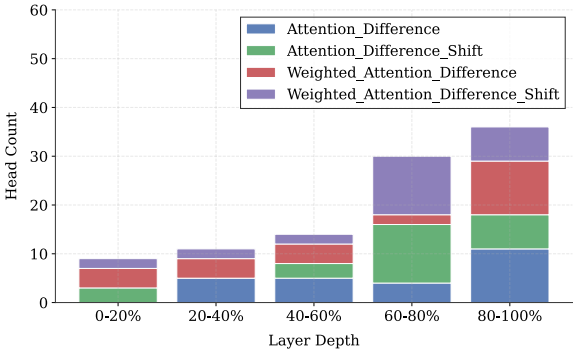


Figure 11: Distribution of the heads identified by the four proposed methods on Qwen2.5-32B

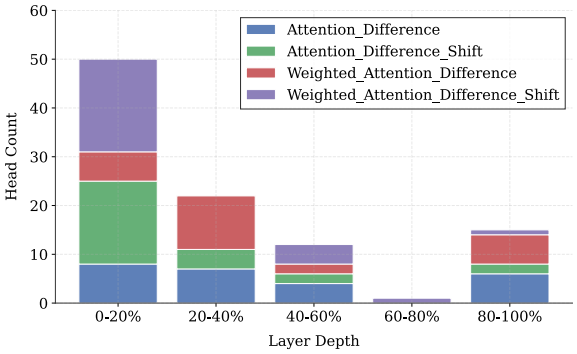


Figure 12: Distribution of the heads identified by the four proposed methods on Llama-3.1-8B.

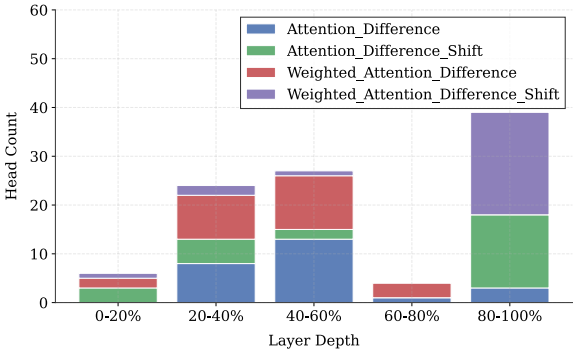


Figure 13: Distribution of the heads identified by the four proposed methods on phi-4

Method	**Qwen2.5-14B**	**Qwen2.5-32B**	**Llama-3.1-8B**	**Phi-4**
Attention_Difference	0.051	0.220	0.000*	0.059
Attention_Difference_Shift	0.009*	0.010*	0.457	0.948
Weighted_Attention_Difference	0.016*	0.807	0.000*	0.943
Weighted_Attention_Difference_Shift	0.586	0.081	1.000	0.219

Table 6: The p-values of the one-tailed paired t-test between the proposed methods and the vanilla LLM under the Anchoring condition (* indicates $p < 0.05$).