Query-Routed Activation Editing with Truth-hierarchical Preference Optimization

Kewei Liao[1,2*], **Tianbo Wang**[1,2*], **Yuqing Ma**[3,2†], **Zhange Zhang**[3,2], **Zhicheng Geng**[3,2], **Xiaowei Zhao**[4], **Jiakai Wang**[4,2], **Xianglong Liu**[1,2]

[1]School of Computer Science and Engineering, Beihang University
[2]State Key Laboratory of Complex & Critical Software Environment
[3]Institute of Artificial Intelligence, Beihang University
[4]Zhongguancun Laboratory
liaokewei@buaa.edu.cn, mayuqing@buaa.edu.cn

Abstract

Hallucination has emerged as a pivotal challenge of Large Language Models (LLMs) that generate plausible yet non-factual content, significantly impeding the trustworthy AI applications in real-world scenarios like medical diagnosis and autonomous driving. Editing the internal activations of LLMs during inference has shown promising effectiveness in mitigating hallucinations with minimal cost. However, previous editing approaches neglect the query-specific inference pathways that require tailored truthful steering vectors, resulting in suboptimal hallucination mitigation. To address these issues, we propose the ***Q**uery-**R**outed **A**ctivation **E**diting (QRAE)* framework, which comprises *Divergence-sensitive Head Routing (DHR)* and *Truth-hierarchical Preference Steering (TPS)*, to fully leverage query-specific semantics for adaptive activation editing. Specifically, DHR is proposed to establish a query-aware head selection criterion, thereby dynamically routing to truth-critical attention heads. Subsequently, TPS introduces a query-specific steering vector calibration policy with the guidance of progressive truth-preferred optimization, enabling precise and adaptive editing for each distinct query. Extensive experiments on the widely recognized TruthfulQA benchmark demonstrate that QRAE outperforms SOTA methods by up to 13.2% in MC1. Meanwhile, QRAE demonstrates strong generalization to out-of-distribution TriviaQA and Natural Questions benchmarks.

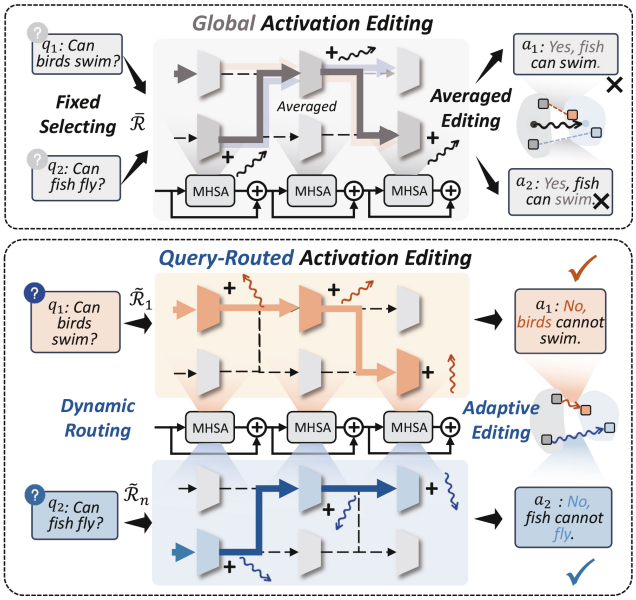


Figure 1: Comparison between previous methods and proposed QRAE on both head selecting and activation editing.

Code — https://github.com/liaokewei/QRAE

1 Introduction

Although Large Language Models (LLMs) have achieved substantial advancements across a wide range of applications (Hu et al. 2024; Bae et al. 2022; Zheng et al. 2025; Huang et al. 2024), they are still susceptible to hallucination (Huang et al. 2025; Rawte, Sheth, and Das 2023), which refers to generating plausible yet nonfactual content. This issue poses significant risks to the trustworthy application of LLMs in real-world scenarios like medical diagnosis (Kim et al. 2025) and autonomous driving (Wang 2024). Numerous studies have explored training-based paradigms like instruction tuning (Wei et al. 2021) and reinforcement learning from human feedback (Ouyang et al. 2022) to effectively mitigate hallucination. However, their practical scalability is constrained by substantial resources and large-scale curated datasets (Casper et al. 2023).

To achieve more efficient hallucination mitigation, inference-time activation editing (Li et al. 2023; Chen et al. 2024) has emerged as a lightweight technique that steers the untruthful activations to a desired truthful space. Specifically, generic editing methods (Li et al. 2023) first identify a fixed set of truth-related attention heads according to averaged scores across the whole dataset. Subsequently, the activations of selected heads will be edited with a pre-computed universal vector to enhance factuality. However, these fixed

*These authors contributed equally.
†Corresponding author

editing strategies lack the flexibility to handle diverse hallucination types and varying input semantics. To address these limitations, recent research has explored more adaptive editing techniques. For instance, ACT (Wang et al. 2025a) generates a set of steering vectors on different hallucination categories and performs adaptive vector aggregation. SADI (Wang, Yang, and Peng 2024) identifies editing locations with a binary mask and then achieves semantic adaptation by scaling the hidden features of the input.

However, although prior methods have achieved promising results, they overlook that hallucinations arise from query-specific inference pathways, which necessitate tailored steering vectors for a distinct query to achieve effective mitigation. Consequently, as illustrated in Figure 1, their performance is markedly limited in both head selecting and activation editing stage: (1) Distinct queries are revealed (Nam et al. 2025) to activate diverse inference pathways within LLM, leading to varying untruthful activations. Therefore, employing the invariant head selection strategy of existing methods fails to capture query-specific untruthful activations, rendering subsequent editing ineffective. (2) Beyond head selection, different query activations exhibit varying factual deviations, revealing diverse steering directions for truthfulness. However, the existing universal editing vector ignores query-specific directional calibration demands, resulting in identical yet imprecise editing.

In this paper, we propose the ***Q**uery-**R**outed **A**ctivation **E**diting (QRAE)* framework, which comprises ***D**ivergence-sensitive **H**ead **R**outing (DHR)* and ***T**ruth-hierarchical **P**reference **S**teering (TPS)*, to fully leverage query-specific semantics for adaptive activation editing. Specifically, DHR establishes a query-variant head selection criterion to adaptively route different queries toward their most relevant inference pathway. Building on the intuition that larger truthful-untruthful sensitivity signifies a greater capacity for improving truthfulness (Li et al. 2023), our criterion adaptively quantifies divergence between the activation distributions of truthful and untruthful response clusters to assess the steering potential of each head. Therefore, the most sensitive heads for the inferred query can be effectively identified, offering promising locations for subsequent editing. Guided by the DHR-selected heads, TPS introduces a query-guided directional calibration policy to adaptively refine the global steering vector for each head. This policy is optimized through coarse-to-fine preference over truth-hierarchical answers, transitioning from untruthful to truthful, and ultimately to the best answer. As a result, TPS derives a tailored editing strategy that maximizes factual responses for each query, enabling highly adaptive and effective steering.

Extensive experiments on the authoritative TruthfulQA benchmark validate the efficacy of QRAE, which achieves a remarkable 85.1% on the primary True*Info metric and significantly surpasses state-of-the-art adaptive methods. Furthermore, QRAE exhibits exceptional generalizability across multiple out-of-distribution datasets, underscoring its robustness and broad application potential.

2 Related Works

2.1 Inference-time Activation Editing

Inference-time activation editing (Li et al. 2023; Chen et al. 2024; Zhang, Yu, and Feng 2024; Wang et al. 2025b) mitigates LLM hallucinations by directly steering activations toward a desired truthful space. Li et al. (2023) first introduced Inference-Time Intervention, a method that selects truth-related attention heads based on linear probe accuracy and computes a universal steering vector from the mean difference between truthful and untruthful activations. TrFr (Chen et al. 2024) enhances this approach by training multiple orthogonal probes per head, combining their directions to form a more comprehensive editing vector. Alternatively, TruthX (Zhang, Yu, and Feng 2024) utilizes an auto-encoder to decouple representations into truthful and semantic spaces, identifying an editing direction from the difference between class means within the truthful space. SEA (Qiu et al. 2024) offers a training-free method that derives editing projections through spectral decomposition of activation covariance matrices from positive and negative demonstrations.

However, these methods use a fixed head selection and a universal editing vector, thus lacking the query-specific calibration needed to handle diverse inference paths and varying factual deviations.

2.2 Adaptive Activation Editing

To enable more granular control, recent works introduce adaptive editing mechanisms that tailor editing to individual inputs. LITO (Bayat et al. 2024) generates multiple candidate responses by applying a universal editing vector at varying strengths and uses a trained classifier to select the most truthful output. ACT (Wang et al. 2025a) generates multiple steering vectors by clustering different hallucination categories and then adaptively modulates editing intensity based on a probe's assessment of the current activation's truthfulness. Similarly, ASTRA (Wang, Wang, and Zhang 2024) determines its editing strength through the projection of a calibrated activation onto a pre-computed harmful direction, ensuring steering is applied only when an input aligns with harmful semantics. In contrast to adapting intensity, SADI (Wang, Yang, and Peng 2024) constructs a dynamic steering vector by identifying critical editing locations and scaling the input's own activations as the editing.

While these methods introduce adaptivity in either editing intensity or vector construction, they do not jointly perform query-specific routing to truth-critical heads and tailored calibration of the steering vector.

3 Methodology

Previous editing methods ignore query variability in both the activation selection and editing stages, leading to imprecise hallucination mitigation. To realize fine-grained activation editing, we introduce the Query-Routed Activation Editing (QRAE) framework. QRAE initially leverages Divergence-sensitive Head Routing (DHR) to establish a query-aware head selection criterion, thereby dynamically routing to truth-critical attention heads. Subsequently, Truth-hierarchical Preference Steering (TPS) introduces a query-

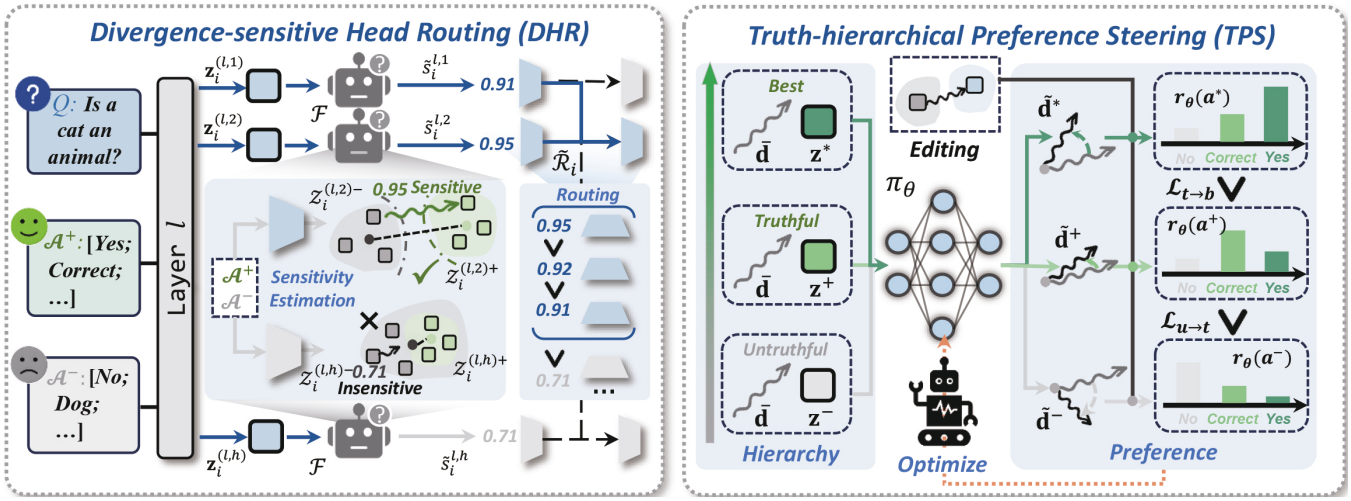


Figure 2: An overview of the QRAE framework. DHR dynamically selects query-relevant attention heads by estimating their truth-untruth sensitivity across activation clusters. TPS then optimizes query-specific steering vectors through coarse-to-fine preference calibration over a truth-hierarchical answer set, enabling precise and adaptive activation editing.

specific steering vector calibration policy with the guidance of progressive truth-preferred optimization, enabling precise and adaptive editing for each distinct query. In this section, we first present the preliminaries of activation editing in Section 3.1, and elaborate on DHR in Section 3.2 and TPS in Section 3.3, respectively.

3.1 Preliminary

Given a Large Language Model (LLM) $\mathcal{M}$ composed of L layers with H attention heads, the model generates an answer $a_i = \mathcal{M}(q_i)$ for an input query q_i. However, the internal activations tend to be untruthful when the query-related knowledge is not properly activated (Li et al. 2023), causing the answer y to be hallucinatory. Therefore, inference-time activation editing technique is devised to directly steer untruthful activations, thereby promoting truthful responses.

Typically, given truthful answers $\mathcal{A}_i^+ = \{a_{ij}^+\}$ and untruthful answers $\mathcal{A}_i^- = \{a_{ik}^-\}$ for query q_i, a triplet dataset $\mathcal{D} = \{(q_i, a_{ij}^+, a_{ik}^-)\}$ is sampled from $\mathcal{A}_i^+$ and $\mathcal{A}_i^-$ for activation editing. Subsequently, a_{ij}^+ and a_{ik}^- are concatenated with query q_i to extract the corresponding activations $\mathbf{z}_{ij}^{(l,h)+}$ and $\mathbf{z}_{ik}^{(l,h)-}$ for each $(l,h) \in \mathcal{H}$, where $\mathcal{H} = [1, L] \times [1, H]$ denotes the Cartesian product of LLM attention heads.

For each head $(l,h) \in \mathcal{H}$, activation editing methods first compute a fixed score $s_{\text{fix}}^{(l,h)}$ (*e.g.* the validation accuracy of head-wise binary classifiers on the validation set $\mathcal{D}^{\text{val}}$ (Li et al. 2023; Chen et al. 2024)) to represent the average importance for all queries:

$$s_{\text{fix}}^{(l,h)} = \frac{1}{2 \cdot |\mathcal{D}^{\text{val}}|} \sum_{\mathcal{D}^{\text{val}}} \Big(\mathbb{I}\big[f^{(l,h)}(\mathbf{z}_{ij}^{(l,h)+}) \geq \tau \mid y_{\mathbf{z}} = 1 \big] + \mathbb{I}\big[f^{(l,h)}(\mathbf{z}_{ik}^{(l,h)-}) < \tau \mid y_{\mathbf{z}} = 0 \big] \Big) \tag{1}$$

where $f^{(l,h)}(\cdot)$ denotes the binary classifier, τ denotes the classification threshold (default to 0.5), $y_{\mathbf{z}}$ denotes the binary truthful label and $\mathbb{I}(\cdot)$ denotes the indicator function. According to the score $s_{\text{fix}}^{(l,h)}$, they select top K heads from $\mathcal{H}$ to form a fixed set $\mathcal{R}_{\text{fix}} \subseteq \mathcal{H}$, which is treated as comprising heads with equal potential for generating untruthful activations for all queries. Consequently, for each selected head $(l,h) \in \mathcal{R}_{\text{fix}}$, a global editing vector $\bar{\mathbf{d}}^{(l,h)}$ (*e.g.* by averaging activation differences between truthful and untruthful samples over $\mathcal{D}$) is constructed as:

$$\bar{\mathbf{d}}^{(l,h)} = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \left[\mathbf{z}_{ij}^{(l,h)+} - \mathbf{z}_{ik}^{(l,h)-} \right]. \tag{2}$$

Subsequently, the global vector $\bar{\mathbf{d}}^{(l,h)}$ is applied to the activations of each head $(l,h) \in \mathcal{R}_{\text{fix}}$ with a fixed strength α:

$$\mathbf{z}'^{(l,h)} = \mathbf{z}^{(l,h)} + \alpha \cdot \bar{\mathbf{d}}^{(l,h)}. \tag{3}$$

However, the previous fixed set of selected heads $\mathcal{R}_{\text{fix}}$ potentially excludes the most informative heads for a given query. In contrast, our DHR in Section 3.2 can dynamically route each inferred query to its most sensitive heads $\tilde{\mathcal{R}}_i$. Additionally, previous editing strategies ignore the query-specific steering direction and apply the globally averaged editing vector $\bar{\mathbf{d}}^{(l,h)}$. Our TPS in Section 3.3 learns to refine the $\bar{\mathbf{d}}^{(l,h)}$ into an optimal query-aware vector $\tilde{\mathbf{d}}_i^{(l,h)}$, thus maximizing the truthfulness of the generated answer.

3.2 Divergence-sensitive Head Routing

To break the previous invariant head-sensitivity criterion, we propose Divergence-sensitive Head Routing (DHR) to adaptively route queries to query-related heads. DHR intuitively leverages the activation divergences between query-related truthful and untruthful answer clusters as a sensitivity metric. It guides the constructed query-specific head estimator to precisely quantify head sensitivity, thereby facilitating effective routing to the appropriate pathways.

Specifically, we first devise a query-specific head estimator $\mathcal{F}(\cdot;\theta_{\mathcal{F}})$, where $\theta_{\mathcal{F}}$ denotes the parameters of $\mathcal{F}$, to dynamically quantify which attention heads are most likely involved in the inference pathway of a given query. In contrast to traditional invariant score $s_{\text{fix}}^{(l,h)}$, our estimator effectively captures the internal hallucinatory patterns of the individual query q_i through the activation $\mathbf{z}_i$, and predicts the adaptive sensitivity score $\tilde{s}_i^{(l,h)}$ for head $(l,h) \in \mathcal{H}$.

$$\tilde{s}_i^{(l,h)} = \mathcal{F}(\mathbf{z}_i^{(l,h)}; \theta_{\mathcal{F}}) \tag{4}$$

This score reflects the head potential for improving factuality within the query-specific inference pathway. Accordingly, by selecting the top K heads with the highest sensitivity score $\tilde{s}_i^{(l,h)}$, we adaptively route the query to its most truthfulness-related heads $\tilde{\mathcal{R}}_i$:

$$\tilde{\mathcal{R}}_i = \{(l,h)_j^i \in \mathcal{H} \mid \mathcal{F}(\mathbf{z}_i^{(l,h)}; \theta_{\mathcal{F}}) \geq \tau_i\}, \tag{5}$$

where τ_i denotes the K-th highest sensitive score across all heads for query q_i. In general, the query-routed heads $\tilde{\mathcal{R}}_i \neq \tilde{\mathcal{R}}_j$ differ across distinct queries, revealing the query-specific nature of factuality-sensitive inference pathways.

Inspired by the discoveries (Guan et al. 2020; Geshkovski et al. 2023) that informative attention head activations tend to aggregate into functional clusters in the latent space, we propose activation distribution divergence as the supervision signal for training the estimator $\mathcal{F}$, which indicates the head's sensitivity to the query-centric truthfulness. Specifically, for each query q_i, we leverage the truthful answers $\mathcal{A}_i^+$ and untruthful answers $\mathcal{A}_i^-$ to extract the truthful activation cluster $\mathcal{Z}_i^{(l,h)+} = \{\mathbf{z}_{ij}^{(l,h)+}\}$ and untruthful activation cluster $\mathcal{Z}_i^{(l,h)-} = \{\mathbf{z}_{ik}^{(l,h)-}\}$ at head $(l,h) \in \mathcal{H}$, respectively. The supervisory divergence $\hat{s}_i^{(l,h)}$ is computed as:

$$\hat{s}_i^{(l,h)} = \frac{\|\mu_i^{(l,h)+} - \mu_i^{(l,h)-}\|_2^2}{\sigma_i^{(l,h)+} + \sigma_i^{(l,h)-} + \epsilon}, \tag{6}$$

where $\mu = \frac{1}{|\mathcal{Z}|}\sum_{\mathbf{z}\in\mathcal{Z}} \mathbf{z}$ denotes the mean vector of an activation cluster, $\sigma = \frac{1}{|\mathcal{Z}|}\sum_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{z}-\mu\|_2^2$ represents its variance, and ϵ is a small positive constant for numerical stability. With the guidance of divergence, we train the estimator $\mathcal{F}$ by minimizing the Mean Squared Error between the predicted scores and the supervisory divergence scores:

$$\mathcal{L}_{DHR} = \frac{1}{|\mathcal{D}| \cdot |\mathcal{H}|} \sum_{\mathcal{D}} \sum_{\mathcal{H}} \|\mathcal{F}(\mathbf{z}_i^{(l,h)}; \theta_{\mathcal{F}}) - \hat{s}_i^{(l,h)}\|_2^2. \tag{7}$$

As a result, our DHR module can route a diverse query to the promising query-specific heads $\tilde{\mathcal{R}}_i$ for subsequent editing.

3.3 Truth-hierarchical Preference Steering

To accommodate the query-specific variability in editing direction, we propose Truth-hierarchical Preference Steering (TPS) to achieve optimal activation steering for each query. TPS incorporates a query-guided directional calibration policy to adaptively refine the global steering vector. It is progressively trained by optimizing the query-centric editing preferences over the constructed truth-stratified answer hierarchy, thereby promoting effective editing on the sensitive heads $\tilde{\mathcal{R}}_i$ routed by DHR.

Initially, we devise the query-guided directional calibration policy π_θ, which perceives the environmental query semantics $\mathbf{z}_i$ [1] and performs optimal directional calibration action of the global vector $\bar{\mathbf{d}}$. To effectively guide policy π_θ toward generating optimal steering vectors that maximize the response factuality, we optimize π_θ with a truth-hierarchical preference, which transitions from untruthful to truthful, and ultimately to the best answer. Specifically:

(1) Basic *untruthful-to-truthful* preference: First, the *untruthful-to-truthful* preference $\mathcal{L}_{u\to t}$ ensures that the optimized query-specific calibration strategy π_θ exhibits a basic truthful-untruthful discriminative advantage. Specifically, policy π_θ generates truthful vectors $\{\tilde{\mathbf{d}}_{ij}^+ | \tilde{\mathbf{d}}_{ij}^+ = \pi_\theta([\bar{\mathbf{d}}, \mathbf{z}_{ij}^+]; \theta)\}$ by calibrating $\bar{\mathbf{d}}$ with the guidance of truthful activations $\{\mathbf{z}_{ij}^+\}$. These vectors are expected to steer LLM towards responding truthfully.

Similarly, policy π_θ leverages the untruthful activations $\{\mathbf{z}_{ik}^-\}$ to refine the global editing vector $\bar{\mathbf{d}}$, generating untruthful vectors $\{\tilde{\mathbf{d}}_{ik}^- | \tilde{\mathbf{d}}_{ik}^- = \pi_\theta([\bar{\mathbf{d}}, \mathbf{z}_{ik}^-]; \theta)\}$ that maximize the probability of hallucinatory responses.

Consequently, the preference $\mathcal{L}_{u\to t}$ constrains the reward $r_\theta(a_{ij}^+, q_i)$ of all truthful answers to be ranked above the reward $r_\theta(a_{ik}^-, q_i)$ of all untruthful answers. The reward is formulated as:

$$r_\theta(a_i, q_i) = \beta \cdot \log \frac{\mathcal{M}(a_i|q_i) \circ \pi_\theta([\bar{\mathbf{d}}, \mathbf{z}_i]; \theta)}{\mathcal{M}(a_i|q_i) \circ \bar{\mathbf{d}}}, \tag{8}$$

where $\mathcal{M}(a_i|q_i) \circ \pi_\theta([\bar{\mathbf{d}}, \mathbf{z}_i]; \theta)$ denotes the LLM's answer distribution after applying the π_θ-calibrated steering vectors to the DHR-selected heads $\tilde{\mathcal{R}}_i$:

$$\mathbf{z}_i' = \mathbf{z}_i + \alpha_i \cdot \pi_\theta([\bar{\mathbf{d}}, \mathbf{z}_i]; \theta), \tag{9}$$

and $\alpha_i = \tilde{s}_i \cdot \alpha_{\max}$ denotes the query-specific editing strength based on the sensitivity score $\tilde{s}_i$. Accordingly, $\mathcal{M}(a_i|q_i) \circ \bar{\mathbf{d}}$ denotes the distribution of LLM with global editing vector $\bar{\mathbf{d}}$. The whole $\mathcal{L}_{u\to t}$ preference can be formulated as:

$$\mathcal{L}_{u\to t} = -\mathbb{E}_{(q_i, \mathcal{A}_i^+, \mathcal{A}_i^-)\sim\mathcal{D}} \Big[\log \phi \Big(\mathbb{E}_{a_{ij}^+ \sim \mathcal{A}_i^+} r_\theta(a_{ij}^+, q_i) - \mathbb{E}_{a_{ik}^- \sim \mathcal{A}_i^-} r_\theta(a_{ik}^-, q_i) \Big) \Big], \tag{10}$$

where ϕ is the logistic function.

(2) Progressive *truthful-to-best* preference: Furthermore, we build the advanced *truthful-to-best* preference $\mathcal{L}_{t\to b}$ to progressively refine the query-aware calibration policy π_θ from merely correct to optimal. To this end, we further perform a fine-grained stratification of the truthful answers and choose the best answer a_i^* for q_i. Within the answer hierarchy $[a_i^*, \mathcal{A}_i^+, \mathcal{A}_i^-]$, a_i^* serves as the most truthful and comprehensive answer, thereby providing the optimal calibration assistance for π_θ to derive the best vector $\tilde{\mathbf{d}}_i^* = \pi_\theta([\bar{\mathbf{d}}, \mathbf{z}_i^*]; \theta)$. This optimal vector should yield

[1] Due to the identical operation for each head (l,h), we omit the (l,h) index for all relevant symbols (*e.g.*, $\mathbf{z}$, $\mathbf{d}$) in the following paper to simplify the notation.

a higher reward $r_\theta(a_i^*, q_i)$ than all truthful vectors, thereby establishing the preference $\mathcal{L}_{t\to b}$:

$$\mathcal{L}_{t\to b} = -\mathbb{E}_{(q_i,a_i^*,\mathcal{A}_i^+)\sim\mathcal{D}^*}\Big[\log\phi\Big(r_\theta(a_i^*,q_i) - \mathbb{E}_{a_{ij}^+\sim\mathcal{A}_i^+} r_\theta(a_{ij}^+,q_i)\Big)\Big], \tag{11}$$

Finally, by combining $\mathcal{L}_{TPS} = \mathcal{L}_{u\to t} + \mathcal{L}_{t\to b}$ for progressive optimization, the calibration policy π_θ can effectively adjust the global vector $\bar{\mathbf{d}}$ to query-optimal steering vectors $\tilde{\mathbf{d}}_i$, thereby maximizing the truthfulness of the response.

4 Experiments

In this section, we present comprehensive experiments on the standard hallucination benchmarks to demonstrate the effectiveness of QRAE in mitigating hallucinations.

4.1 Experimental Setup

Benchmarks and Metrics To ensure a comprehensive evaluation of model truthfulness, we conduct comparative experiments with other methods on the TruthfulQA benchmark (Lin, Hilton, and Evans 2022) and assess the generalization ability of our approach on TriviaQA (Joshi et al. 2017) and Natural Questions (Kwiatkowski et al. 2019). For datasets that do not inherently contain hierarchical or multi-level answers, we construct such structures by following the TruthfulQA protocol, which can be readily generalized to other activation editing datasets. In cases where datasets provide a single standardized best answer and the queries are clearly defined without requiring additional human adjudication, we instruct GPT-4 to generate several diversified yet faithful variants. Following established protocols (Li et al. 2023), we assess performance on the open-ended generation task using the True*Info rate, which measures both truthfulness and informativeness. For the multiple-choice task, performance is evaluated with the MC1, MC2 and MC3 accuracy metrics.

Baseline and Comparative Methods We adopt several open-source LLMs as baselines, with LLaMA-3-8B-Instruct (Meta 2024) selected as the primary model for our experimental evaluation.

We first compare the proposed QRAE with various editing-based methods, including fixed editing methods such as ITI (Li et al. 2023), TrFr (Chen et al. 2024), TruthX (Zhang, Yu, and Feng 2024), SEA (Qiu et al. 2024), and HPR (Pham and Nguyen 2024), as well as adaptive editing methods such as LITO (Bayat et al. 2024), ACT (Wang et al. 2025a), and SADI (Wang, Yang, and Peng 2024). We also consider other effective methods that enhance various alignment capabilities for comparison, including Supervised Fine-Tuning (SFT) (Li et al. 2023), Few-shot Prompting (FSP) (Bai et al. 2022), and decoding-based approaches like DoLa (Chuang et al. 2023) and SH2 (Kai et al. 2024).

Implementation Details We adhere to the experimental setup outlined in (Li et al. 2023) and apply 2-fold validation for all experiments, thereby ensuring fair comparisons. Unless noted, the number of edited heads K is set to 48. It is further noticed that the maximum editing strength $\alpha_{\max}$ is set to 20, which serves as a universal setting since the learnable TPS scales vectors to compensate. Trainable modules are optimized with the Adam optimizer at a learning rate of 1×10^{-3}. All experiments run on NVIDIA A100 GPUs.

4.2 Experimental Results

As shown in Table 1, our proposed QRAE framework consistently outperforms all existing methods on the TruthfulQA benchmark, demonstrating strong effectiveness in both open-ended and multiple-choice settings. For open-ended generation, QRAE achieves a True*Info score of 85.1%, surpassing the baseline of 62.0% by 23.1 points. It also attains 63.3% accuracy on MC1, a 13.2% improvement over the next-best method. These results underscore QRAE's ability to generate truthful and informative responses, attributed to its Divergence-sensitive Head Routing and Truth-hierarchical Preference Steering, which enable adaptive path selection and query-specific vector calibration. Compared with other editing-based methods, QRAE not only enhances response truthfulness but also maximizes information content, reaching an Info metric of 95%. Notably, QRAE does not exceed some methods on this metric because TruthfulQA includes questions requiring non-informative answers such as *"I have no comment"* (56/817). These responses in our method (36/817 questions) contribute to the slightly lower Info metric, yet are closely aligned with the distribution of answers in the dataset.

Besides the powerful LLaMA-3-Instruct-8B, we also validate eight other sophisticated LLMs varying in architecture and parameter size. Figure 3 shows that we can effectively enhance the truthfulness of all models, yielding average improvements of 24.4% in True*Info score.

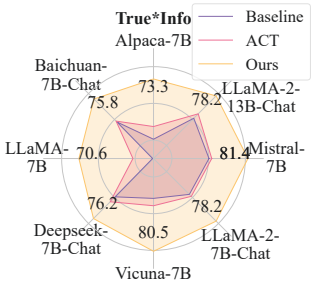


Figure 3: Performance of QRAE across various LLMs.

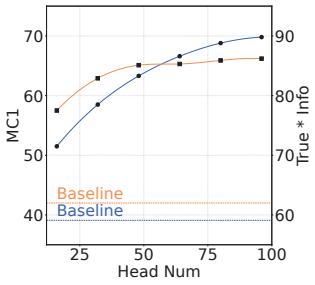


Figure 4: MC1 and True*Info with K heads.

4.3 Ablation Study

To validate the individual contributions of our proposed modules, we conduct an ablation study with results presented in Table 2. The baseline model without any editing achieves a True*Info score of 62.0%. Integrating only the DHR module improves the True*Info score to 74.2%, which demonstrates the effectiveness of dynamically routing queries to relevant attention heads. Applying only the TPS module yields a more substantial improvement, increasing the score by 21.3% and underscoring the critical role of

Methods	Open-ended Generation			Multiple-Choice		
	True*Info (↑)	True (↑)	Info (↑)	MC1 (↑)	MC2 (↑)	MC3 (↑)
Baseline	62.0	69.5	89.2	39.1	58.6	29.5
Supervised Fine-tuning	69.5	71.2	97.6	39.3	56.6	30.6
Few-shot Prompting	66.4	67.4	98.4	41.4	59.2	29.6
Decoding-based Methods						
DoLa *(ICLR'24)*	71.8	73.2	**98.0**	40.6	59.3	31.8
SH2 *(EMNLP'24)*	62.3	71.9	86.7	32.2	56.5	31.9
Editing-based Methods						
ITI *(NeurIPS'23)*	69.0	79.8	86.4	41.1	61.1	31.7
TrFr *(AAAI'24)*	73.2	82.0	89.3	41.5	60.0	30.8
TruthX *(ACL'24)*	64.9	71.8	90.3	42.8	61.2	32.2
LITO *(ACL'24)*	52.6	84.6	62.3	40.4	58.3	29.6
SEA *(NeurIPS'24)*	72.3	81.8	88.4	42.8	61.1	33.3
ACT *(WWW'25)*	72.6	79.6	91.2	42.2	62.1	32.1
SADI *(ICLR'25)*	75.2	83.1	90.5	43.1	62.6	32.4
HPR *(EMNLP'24)*	78.5	85.1	92.2	50.1	68.2	42.4
Ours	**85.1**(↑**6.6%**)	**89.6**(↑**4.5%**)	95	**63.3**(↑**13.2%**)	**77.9**(↑**9.7%**)	**55**(↑**12.6%**)

Table 1: Comparison of QRAE with SOTA methods on the TruthfulQA Benchmark using LLaMA-3-Instruct-8B. The best results are in **bold**. Each numerical result is reported under multiple rounds.

DHR	TPS	True*Info	True	MC1
Baseline		62.0	69.5	39.1
✓		74.2(↑12.2%)	82.3(↑12.8%)	46.3(↑7.2%)
	✓	83.3(↑21.3%)	89.1(↑19.6%)	61.3(↑22.2%)
✓	✓	**85.1**(↑**23.1%**)	**89.6**(↑**20.1%**)	**63.3**(↑**24.2%**)

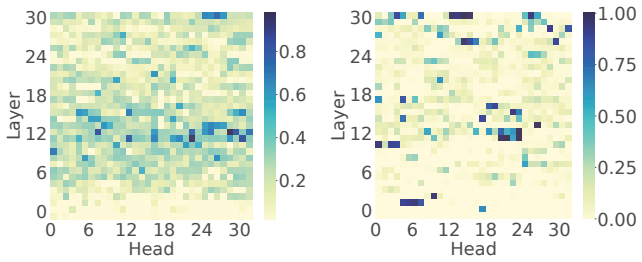
Table 2: The ablation study of two modules in QRAE.

tailored directional calibration. The complete QRAE framework, combining both DHR and TPS, achieves the highest score of 85.1%. This result confirms that our two modules are complementary, with DHR providing the optimal locations for the precise vector calibration performed by TPS.

4.4 Deep Analysis

Analysis of Edited Heads (K) We analyze the impact of the number of edited heads K on QRAE performance, as shown in Figure 4. As K increases from 16 to 96, both the True*Info score and MC1 accuracy steadily improve, with the largest gain observed when K increases from 16 to 48, during which True*Info rises from 77.5% to 85.1% and MC1 from 51.5% to 63.3%. Beyond this range, the improvements become marginal, indicating that the earlier-ranked heads contribute most to factuality enhancement. These results confirm the effectiveness of the DHR module in prioritizing truth-critical heads.

Analysis of Selection Criterion To verify the effectiveness of DHR in identifying truth-critical heads, we compare it with several representative strategies, including Random selection, Bias-based scoring, SADI, and the probe-accuracy method from ITI. All methods utilize the same global editing vector to ensure fairness. As shown in Table 3, DHR achieves the highest performance across all metrics, increasing the True*Info score to 74.2% and MC1 accuracy to 46.3%, which corresponds to improvements of 12.2% and 7.2% over the baseline, respectively. Figure 5 further illustrates that, in contrast to ITI, which primarily clusters heads in mid-layers such as layers 10 to 18, DHR selects a sparser and more widely distributed set across the network. These results demonstrate that DHR is capable of capturing query-specific, truth-sensitive heads beyond what can be achieved by global averaging strategies.



(a) ITI probe accuracy scores. (b) DHR query-specific scores.

Figure 5: Visualization of head selection methods.

Analysis of the Hierarchical Calibration Strategy We analyze the components of the TPS module, with results shown in Table 4. Using DHR alone (without TPS calibration) yields a True*Info score of 74.2%. Introducing only the untruthful-to-truthful loss ($\mathcal{L}_{u\to t}$) boosts performance to 81.3%, highlighting its key role in establishing truthful preference. In contrast, using only the truthful-to-best loss ($\mathcal{L}_{t\to b}$) yields smaller gains, suggesting it mainly refines the truthful direction. Combining both losses, the full TPS module achieves the highest score of 85.1%, demonstrating

Methods	True*Info	True	MC1
Baseline	62.0	69.5	39.1
Random	65.8	71.9	40.2
ITI	69.0	79.8	41.1
Bias	52.2	65.9	39.5
SADI	66.2	73.1	40.8
Ours	74.2(↑**12.2%**)	82.3(↑**12.8%**)	46.3(↑**7.2%**)

Table 3: Comparison of different head selection methods.

Methods	True*Info	True	MC1
w/o TPS	74.2	82.3	46.3
w/ $\mathcal{L}_{u\to t}$	81.3(↑7.1%)	88.3(↑6.0%)	59.7(↑13.4%)
w/ $\mathcal{L}_{t\to b}$	78.5(↑4.3%)	86.2(↑3.9%)	53.3(↑7.0%)
w/ TPS	**85.1**(↑**10.9%**)	**89.6**(↑**7.3%**)	**63.3**(↑**17.0%**)

Table 4: Analysis of different preferences in TPS.

the effectiveness of modeling truth-hierarchical preference. Rather than relying on a single signal, TPS adopts a two-stage strategy in which truthfulness is aligned progressively from coarse to fine, allowing the steering vector to be refined toward accurate and informative answers. These results confirm the importance of hierarchical preference modeling for precise, query-specific editing.

Analysis of Training Data Size To assess data efficiency, we evaluate QRAE under varying training data sizes, as shown in Figure 6. With only 10% of the data, QRAE achieves a strong MC1 score of 48.7%, and performance continues to improve with more data, reaching 58.9% at 50%. These results demonstrate that QRAE performs well in low-data regimes while scaling effectively, as additional data helps DHR better identify relevant heads and allows TPS to refine the editing vector more precisely. This highlights both the framework's strong performance in low-data regimes and its capacity to achieve higher precision as more data becomes available.

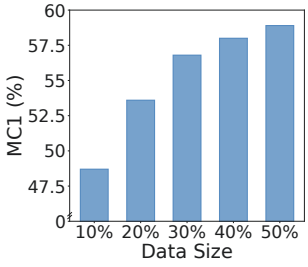


Figure 6: MC1 performance across data size.

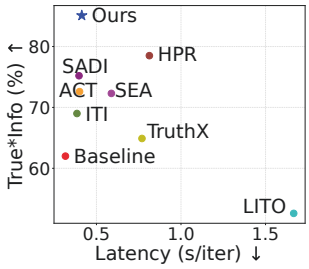


Figure 7: Comparison of Inference Computation.

4.5 Inference Computation

We compare the inference efficiency of QRAE with other editing methods in Figure 7, where Latency measures the time per iteration (s/iter). QRAE achieves the highest True*Info score among all methods while introducing only a modest increase in inference time over the baseline. It remains more efficient than several adaptive methods such as SADI and ACT. Given the substantial gains in factuality, this moderate overhead is acceptable, demonstrating that QRAE effectively balances performance and efficiency.

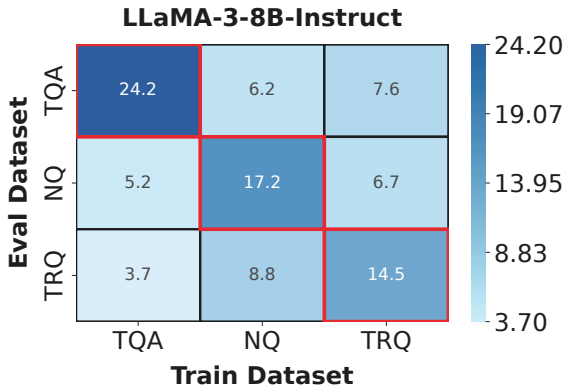


Figure 8: Cross-dataset generalization of QRAE.

4.6 Generalizability

To assess the generalizability of QRAE, we perform a cross-dataset evaluation where the DHR and TPS modules are trained on one benchmark and applied to others without further tuning. As shown in Figure 8, QRAE achieves consistent improvements over the baseline across all train-test combinations. Each cell reports the absolute percentage point gain in the MC1 metric. The diagonal entries reflect in-domain effectiveness, with the highest improvement of 24.2% observed when both training and evaluation are conducted with TruthfulQA. Importantly, strong cross-dataset transferability is also evident. For example, models trained on Natural Questions yield 8.8% and 6.2% gains on TriviaQA and TruthfulQA, respectively. These results demonstrate that QRAE does not rely on dataset-specific cues. Instead, its query-adaptive editing strategy captures generalizable patterns of factual reasoning, enabling effective hallucination mitigation even on unseen distributions.

5 Conclusion and Future Work

In this work, we addressed the challenge of query-agnostic editing for LLM hallucination by proposing Query-Routed Activation Editing (QRAE), a novel framework for adaptive activation editing. Our framework uniquely combines divergence-sensitive head routing with truth-hierarchical preference steering to achieve precise, context-aware editing. Extensive experiments confirmed that QRAE significantly outperforms existing methods on several challenging benchmarks. While effective, QRAE relies on a truth hierarchy during TPS calibration, which may limit its use for queries that cannot be verified for correctness. Future work will explore dynamic preference modeling to adapt the hierarchy based on query intent or context.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2023YFC2506800), the Fundamental Research Funds for the Central Universities and Young Elite Scientists Sponsorship Program of the Beijing High Innovation Plan.

References

Bae, S.; Kwak, D.; Kim, S.; Ham, D.; Kang, S.; Lee, S.-W.; and Park, W. 2022. Building a role specified open-domain dialogue system leveraging large-scale language models. *arXiv preprint arXiv:2205.00176*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bayat, F. F.; Liu, X.; Jagadish, H.; and Wang, L. 2024. Enhanced Language Model Truthfulness with Learnable Intervention and Uncertainty Expression. In *Findings of the Association for Computational Linguistics ACL 2024*, 12388–12400.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Chen, Z.; Sun, X.; Jiao, X.; Lian, F.; Kang, Z.; Wang, D.; and Xu, C. 2024. Truth forest: Toward multi-scale truthfulness in large language models through intervention without tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 20967–20974.

Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J. R.; and He, P. 2023. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Geshkovski, B.; Letrouit, C.; Polyanskiy, Y.; and Rigollet, P. 2023. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36: 57026–57037.

Guan, Y.; Leng, J.; Li, C.; Chen, Q.; and Guo, M. 2020. How Far Does BERT Look At: Distance-based Clustering and Analysis of BERT's Attention. In *Proceedings of the 28th International Conference on Computational Linguistics*, 3853–3860.

Hu, J.; Gao, H.; Yuan, Q.; and Shi, G. 2024. Dynamic content generation in large language models with real-time constraints.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Huang, W.; Zheng, X.; Ma, X.; Qin, H.; Lv, C.; Chen, H.; Luo, J.; Qi, X.; Liu, X.; and Magno, M. 2024. An empirical study of llama3 quantization: From llms to mllms. *Visual Intelligence*, 2(1): 36.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1601–1611.

Kai, J.; Zhang, T.; Hu, H.; and Lin, Z. 2024. SH2: Self-Highlighted Hesitation Helps You Decode More Truthfully. In *Findings of the Association for Computational Linguistics: EMNLP 2024"*, 4514–4530.

Kim, Y.; Jeong, H.; Chen, S.; Li, S. S.; Lu, M.; Alhamoud, K.; Mun, J.; Grau, C.; Jung, M.; Gameiro, R.; et al. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36: 41451–41530.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. Accessed: 2024-08-14.

Nam, A.; Conklin, H.; Yang, Y.; Griffiths, T.; Cohen, J.; and Leslie, S.-J. 2025. Causal Head Gating: A Framework for Interpreting Roles of Attention Heads in Transformers. *arXiv preprint arXiv:2505.13737*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pham, V.-C.; and Nguyen, T. 2024. Householder Pseudo-Rotation: A Novel Approach to Activation Editing in LLMs with Direction-Magnitude Perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13737–13751.

Qiu, Y.; Zhao, Z.; Ziser, Y.; Korhonen, A.; Ponti, E. M.; and Cohen, S. B. 2024. Spectral Editing of Activations for Large Language Model Alignment. *Advances in Neural Information Processing Systems*.

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

Wang, H.; Wang, G.; and Zhang, H. 2024. Steering Away from Harm: An Adaptive Approach to Defending Vision Language Model Against Jailbreaks. *arXiv preprint arXiv:2411.16721*.

Wang, J. 2024. Hallucination Reduction and Optimization for Large Language Model-Based Autonomous Driving. *Symmetry*, 16(9): 1196.

Wang, T.; Jiao, X.; Zhu, Y.; Chen, Z.; He, Y.; Chu, X.; Gao, J.; Wang, Y.; and Ma, L. 2025a. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, 2562–2578.

Wang, T.; Ma, Y.; Liao, K.; Yang, C.; Zhang, Z.; Wang, J.; and Liu, X. 2025b. Token-Aware Editing of Internal Activations for Large Language Model Alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 9482–9520.

Wang, W.; Yang, J.; and Peng, W. 2024. Semantics-adaptive activation intervention for llms via dynamic steering vectors. *arXiv preprint arXiv:2410.12299*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Zhang, S.; Yu, T.; and Feng, Y. 2024. Truthx: Alleviating hallucinations by editing large language models in truthful space. *arXiv preprint arXiv:2402.17811*.

Zheng, X.; Li, Y.; Chu, H.; Feng, Y.; Ma, X.; Luo, J.; Guo, J.; Qin, H.; Magno, M.; and Liu, X. 2025. An empirical study of qwen3 quantization. *arXiv preprint arXiv:2505.02214*.