QuantEL: Quantization-Informed Activation Editing Localization for Hallucination Mitigation in Quantized LLMs

Tianbo Wang[1,3] **Lingyan Meng**[1] **Yuqing Ma**[2,3*] **Zhange Zhang**[2,3]
Jian Yang[1,3] **Jinyang Guo**[2,3] **Aishan Liu**[1,3] **Xianglong Liu**[1,3]
School of Computer Science and Engineering, Beihang University[1]
Institute of Artificial Intelligence, Beihang University[2]
State Key Laboratory of Complex & Critical Software Environment[3]
{tianbowang,mayuqing}@buaa.edu.cn

Abstract

Despite the practical benefits of quantized large language models (LLMs) for resource-constrained deployment, they remain vulnerable to hallucinations due to quantization-induced information loss, posing serious reliability risks in real-world applications. Activation editing provides an effective and efficient solution for mitigating hallucinations. However, existing methods identify editable heads primarily based on probe accuracy, failing to account for quantization-induced degradation and thus struggling to prioritize the repair-critical heads in quantized LLMs. To address these limitations, we propose Quantization-informed Editing Localization (QuantEL) by integrating Degradation-Aware Head Selection (DHS) and Quant-Driven Degradation Compensation (QDC) to promote repair-critical head identification. Specifically, DHS estimates head-wise degradation from the discrepancy between full-precision and quantized probe scores to reassess the head editing value for selection. Subsequently, QDC further refines the estimation by modeling quantization-specific discrepancies from both the activation and weight perspectives. Extensive experiments on TruthfulQA demonstrate the efficacy of QuantEL, achieving up to 26.1% improvement on the primary True*Info metric with minimal overhead.

1 Introduction

Although large language models (LLMs) (Meta, 2024; Bi et al., 2024; Zhang et al., 2025) have profoundly transformed productivity and lifestyle, their substantial memory footprint and computational overhead make them difficult to deploy in resource-constrained settings. By mapping full-precision floating-point parameters into low-precision discrete representations, LLM quantization (Frantar et al., 2022; Dettmers et al., 2022) significantly reduces storage and computation costs, thereby enabling practical deployment in edge applications such as autonomous driving (Huang et al., 2025; Mao et al., 2023).

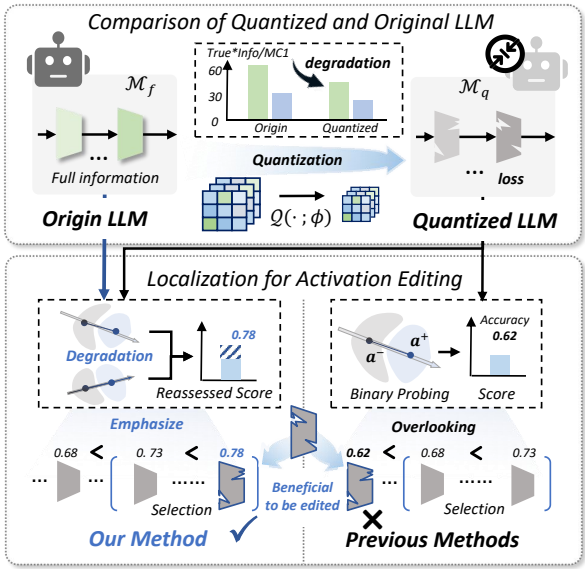

Figure 1: Comparisons between the original and quantized LLMs, and between previous probe-only head selection methods and our QuantEL.

However, as illustrated in Figure 1, quantized LLMs are more vulnerable to hallucinations due to quantization-induced information loss (Wang et al., 2025a), often generating erroneous responses that significantly undermine the safety and reliability of real-world applications. Although training-based alignment methods, such as supervised fine-tuning (SFT) (Wang et al., 2022), have achieved considerable success for full-precision models, their prohibitive training costs make them less suitable for quantized models under low-resource constraints.

To mitigate hallucinations more efficiently, recent studies have explored directly steering LLM behavior at inference time. Several works (Li et al., 2024; Bayat et al., 2024) have identified and validated the interpretable correlation between internal activation space and truthfulness by training probes

*Corresponding author

on truthful and untruthful samples. Therefore, they directly edited the internal activations of attention heads during inference to reduce hallucinated content (Li et al., 2024). This reduces the need for extensive training data and resources, thereby holding substantial application value.

However, as shown in Figure 1, most existing methods (Li et al., 2024; Chen et al., 2024; Wang et al., 2025b, 2026) identify editable heads primarily based on probe accuracy, overlooking the unique necessity to compensate for quantization-induced degradation. As a result, they fail to pinpoint the most valuable heads for editing in quantized LLMs, limiting their ability to mitigate hallucination. Specifically, quantization-induced degradation (Wang et al., 2025a) characterizes the discrepancy between head behaviors before and after quantization. It reflects the extent to which critical information is lost during quantization, thereby indicating the potential benefit that should be prioritized for editing. Moreover, different quantization strategies induce information loss through distinct mechanisms, yielding heterogeneous patterns of head degradation and recovery needs. This highlights the necessity of a quantization-aware compensation strategy for degradation estimation that explicitly adapts to the underlying mechanism.

In this paper, we propose **Quant**ization-informed **E**diting **L**ocalization (**QuantEL**) framework, which comprises ***D**egradation-Aware **H**ead **S**election (**DHS**)* and ***Q**uant-Driven **D**egradation **C**ompensation (**QDC**)*, to identify the attention heads with the highest editing value in quantized LLMs. Given that probe accuracy captures the information effectiveness of attention heads (Li et al., 2024), DHS first estimates the quantization-induced degradation through comparing the accuracies between the full-precision and quantized LLM. By incorporating the degradation into editing value estimation, DHS reliably identifies the most beneficial editing candidates for quantized LLMs. To promote fine-grained degradation estimation, QDC further derives strategy-sensitive compensation signals by characterizing the discrepancy between the quantized and full-precision models from both the activation and weight perspectives. As a result, QDC enables adaptive localization of high-value heads under different quantization strategies, improving the performance of activation editing.

Extensive experiments on TruthfulQA across three LLMs and three quantization strategies demonstrate that QuantEL effectively improves the factuality of both discriminative and generative answers with minimal cost, achieving up to 26.1% improvement on the primary True*Info metric. In summary:

- We propose QuantEL, a localization framework to identify the high-value attention heads for activation editing in quantized LLMs.
- We introduce DHS, which explicitly estimates quantization-induced degradation to promote valuable head selection.
- We introduce QDC, which further optimizes degradation estimation with strategy-sensitive compensation to improve editing localization.
- Extensive experiments on TruthfulQA across three LLMs and three quantization strategies demonstrate the effectiveness of QuantEL.

2 Related Works

LLM Quantization LLM quantization (Lin et al., 2024; Frantar et al., 2022; Xiao et al., 2023) reduces memory and computational overhead by mapping full-precision parameters into low-precision representations, thereby improving the deployment efficiency of LLM in resource-constrained settings. For example, GPTQ (Frantar et al., 2022) performs post-training quantization through block-wise weight reconstruction, enabling accurate low-bit compression with limited calibration data. AWQ (Lin et al., 2024) further improves quantization quality by preserving salient weights according to activation-aware importance, thereby better maintaining model performance under aggressive compression. However, study (Wang et al., 2025a) reveals that quantization inevitably introduces information loss, which may distort internal activation and induce head-wise degradation in quantized LLMs.

Activation Editing Positing that the interpretable internal structures related to truthfulness can be utilized to reduce hallucinated content (Li et al., 2024), activation editing (Qiu et al., 2024; Zhang et al., 2024; Liao et al., 2025, 2026) becomes an efficient technique for hallucination mitigation by directly editing the hallucinatory internal activation during inference. For example, Li et al. (Li et al., 2024) initially introduced the practice of activation editing and demonstrated its capability to enhance the truthfulness of responses. Chen et al.

(Chen et al., 2024) then utilized multiple orthogonal probes to model truthful directions across various dimensions. However, these methods identify editable heads primarily based on probe accuracy and overlook the unique necessity to compensate for quantization-induced degradation, leading to suboptimal performance in quantized LLMs.

3 Methodology

To prioritize the repair-critical heads in quantized LLMs overlooked by previous probe-only head selection strategies, we introduce the Quantization-informed Editing Localization (QuantEL) framework. As shown in Figure 2, QuantEL first leverages Degradation-Aware Head Selection (DHS) to incorporate quantization-induced degradation into head value estimation, and employs Quant-Driven Degradation Compensation (QDC) to further optimize the estimation according to diverse quantization strategies, thereby effectively localizing the heads with the highest editing value. In this section, we first review the process of activation editing in Section 3.1, and elaborate on DHS in Section 3.2 and QDC in Section 3.3.

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 information is not properly activated (Li et al., 2024), 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 the truthful answer a_i^+ and the untruthful answer a_i^- for query q_i, a triplet dataset $\mathbf{D} = \{(q_i, a_i^+, a_i^-)\}$ is first constructed for activation editing. Subsequently, a_i^+ and a_i^- are concatenated with query q_i to extract the corresponding activations $\mathbf{z}_i^{(l,h)+}$ and $\mathbf{z}_i^{(l,h)-}$ for each head $(l, h) \in \mathbf{H}$, where $\mathbf{H} = [1, L] \times [1, H]$ denotes the Cartesian product of LLM attention heads.

For each head $(l, h) \in \mathbf{H}$, activation editing methods first compute the score $s^{(l,h)}$ (*e.g.* the validation accuracy of head-wise binary probes on the validation set $\mathbf{D}^{\text{val}}$ (Li et al., 2024)) to represent the head-wise editing value:

$$s^{(l,h)} = \frac{1}{2 \cdot |\mathbf{D}^{\text{val}}|} \sum_{\mathbf{D}^{\text{val}}} \Big(\mathbb{I}\big[f^{(l,h)}(\mathbf{z}_i^{(l,h)+}) \geq \tau \mid y_{\mathbf{z}} = 1 \big] + \mathbb{I}\big[f^{(l,h)}(\mathbf{z}_i^{(l,h)-}) < \tau \mid y_{\mathbf{z}} = 0 \big] \Big) \quad (1)$$

where $f^{(l,h)}(\cdot)$ denotes the binary classification probe, 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 $s^{(l,h)}$, they select the top K heads from $\mathbf{H}$ to form the editing set $\mathbf{R} \subseteq \mathbf{H}$, which is treated as comprising heads with the highest potential for generating untruthful activations. Consequently, for each selected head $(l, h) \in \mathbf{R}$, an editing vector $\bar{\mathbf{d}}^{(l,h)}$ (*e.g.* by averaging activation differences between truthful and untruthful samples over $\mathbf{D}$) is constructed and applied to the activations of each head $(l, h) \in \mathbf{R}$ to optimize the behavior of $\mathcal{M}$:

$$\mathbf{z}'^{(l,h)} = \mathbf{z}^{(l,h)} + \alpha \cdot \bar{\mathbf{d}}^{(l,h)}, \quad (l, h) \in \mathbf{R}$$
$$\bar{\mathbf{d}}^{(l,h)} = \frac{1}{|\mathbf{D}|} \sum_{\mathbf{D}} \left[\mathbf{z}_i^{(l,h)+} - \mathbf{z}_i^{(l,h)-} \right]. \quad (2)$$

where α denotes the editing strength. However, for a quantized LLM $\mathcal{M}_q = \mathcal{Q}(\mathcal{M}_f; \phi)$, where $\mathcal{M}_f$ denotes the full-precision LLM and $\mathcal{Q}(\cdot; \phi)$ denotes the quantization operation under a quantization strategy ϕ (*e.g.* GPTQ), the previous set of selected heads $\mathbf{R}$ potentially excludes the most valuable heads for mitigating hallucinations arising from quantization-induced information loss. In contrast, our DHS in Section 3.2 and QDC in Section 3.3 collaborate to precisely reassess head-wise intervention priority under quantization, thereby steering activation editing toward the heads with the greatest recovery potential.

3.2 Degradation-Aware Head Selection

We first propose Degradation-Aware Head Selection (DHS) to reassess the editing value under quantization, thereby guiding more precise selection of beneficial heads in quantized LLMs. Rather than relying solely on the post-quantization probe accuracy, DHS explicitly incorporates quantization-induced degradation into head estimation. As a result, DHS prioritizes the valuable heads that remain editable after quantization while also exhibiting recoverable degradation, yielding a more effective editing set for quantized LLMs.

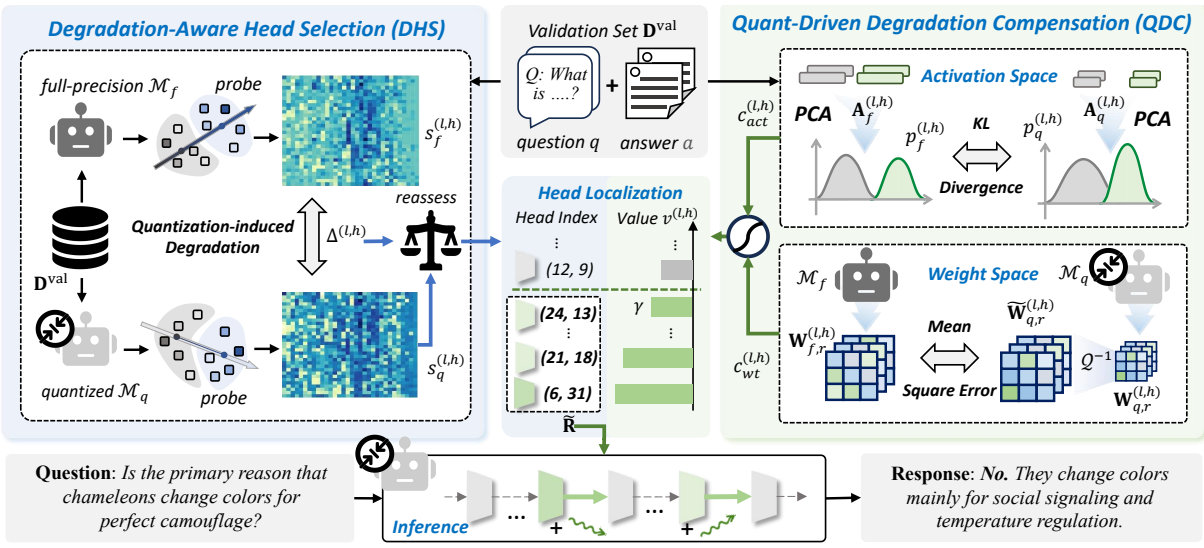

Figure 2: Overview of QuantEL. DHS estimates quantization-induced degradation from the discrepancy between full-precision and quantized probe scores to reassess head editing value and select repair-critical heads. QDC further refines the selected heads by modeling activation-space and weight-space discrepancy, yielding the final editing set for activation editing in quantized LLMs.

Quantization-induced Degradation Estimation Given that probe accuracy captures the extent to which a head preserves truth-discriminative structure (Li et al., 2024), DHS intuitively leverages the quantization-induced drop in probe score as a proxy for estimating the degree of head-wise degradation. Specifically, under the same probing protocol in Eq. 1, we obtain the corresponding probe scores for the full-precision model and its quantized counterpart on the same validation set, denoted by $s_f^{(l,h)}$ and $s_q^{(l,h)}$, respectively. We then define the quantization-induced information degradation of head (l, h) as the normalized positive discrepancy between $s_f^{(l,h)}$ and $s_q^{(l,h)}$:

$$\Delta^{(l,h)} = \frac{\left[s_f^{(l,h)} - s_q^{(l,h)}\right]_+}{s_q^{(l,h)} + \epsilon}, \tag{3}$$

where $\epsilon > 0$ is a small constant for numerical stability and $[x]_+ = \max(x, 0)$ denotes the positive-part operator. Intuitively, a larger $\Delta^{(l,h)}$ indicates more severe functional degradation by quantization, and greater potential benefit from subsequent editing. It is noticed that the positive-part operator $[\cdot]_+$ excludes heads whose probe scores are preserved or even improved after quantization. This is because such heads are unlikely to represent the primary repair bottlenecks and may instead indicate compensatory information reallocation.

Value-reassessed Head Selection Based on the estimated information degradation, DHS reassesses the head editing value for quantized LLMs by jointly capturing post-quantization editability and quantization-induced repair demand, *i.e.*, whether a head still retains usable truthful structure in the quantized model while exhibiting non-negligible information loss. Specifically, the value score $v^{(l,h)}$ of each head is defined as:

$$v^{(l,h)} = s_q^{(l,h)} + \lambda \Delta^{(l,h)}, \tag{4}$$

where λ is a trade-off coefficient. We then rank all heads in $\mathbf{H}$ in descending order according to $v^{(l,h)}$ and select the top K heads as the final editing candidate set $\tilde{\mathbf{R}}$:

$$\tilde{\mathbf{R}} = \{(l, h) \in \mathcal{H} \mid v^{(l,h)} \geq \gamma\}, \tag{5}$$

where γ denotes the K-th highest value score across all heads. In this way, $\tilde{\mathbf{R}}$ serves as the most valuable editing heads in quantized LLMs, thereby improving the effectiveness of subsequent activation editing.

3.3 Quant-Driven Degradation Compensation

Different quantization strategies ϕ distort head behaviors via heterogeneous mechanisms, yielding distinct information degradation from the original LLM (Jin et al., 2024). This motivates fine-grained compensation of degradation estimation that accounts for the underlying quantization mechanism.

To this end, we propose Quant-Driven Degradation Compensation (QDC), which further assesses quantization-induced degradation from internal activations and model weights. By incorporating these two core quantization targets (Wan et al., 2023), QDC provides a more comprehensive compensation for degradation estimates based solely on probe accuracy, thereby improving the accuracy of valuable-head identification.

Activation-space Compensation QDC first characterizes activation-space loss by measuring the distributional discrepancy between the full-precision and quantized activations of each selected head after one-dimensional PCA projection. Taking the full-precision model as an example, for each head (l, h), we apply PCA to its activations on the validation set and retain only the first principal component, yielding a one-dimensional activation set denoted by $\mathbf{A}_f^{(l,h)}$. We then model the resulting projected activations of the full-precision model as a Gaussian distribution:

$$p_f^{(l,h)} = \mathcal{N}\Big(\mu(\mathbf{A}_f^{(l,h)}), \sigma^2(\mathbf{A}_f^{(l,h)})\Big). \quad (6)$$

where $\mu(\cdot)$ and $\sigma^2(\cdot)$ denote the empirical mean and variance of the projected activation set, respectively. The quantized activation distribution $p_q^{(l,h)}$ is obtained analogously by applying the same procedure to the corresponding activations from the quantized model $\hat{\mathcal{M}}$ under the quantization strategy ϕ. We then define the activation-space compensation term as the KL divergence between the quantized and full-precision activation distributions:

$$c_{\text{act}}^{(l,h)} = \mathcal{D}_{\text{KL}}\Big(p_q^{(l,h)} \,\Big\|\, p_f^{(l,h)}\Big). \quad (7)$$

Weight-space Compensation In addition to the activation shift, quantization also perturbs the head-specific projection weights, providing a complementary view for assessing head-wise degradation. Specifically, for each head (l, h) and each projection type $r \in \{Q, K, V\}$, let $\mathbf{W}_{f,r}^{(l,h)}$ and $\mathbf{W}_{q,r}^{(l,h)}$ denote the original weight and quantized counterpart under strategy ϕ. We first dequantize the quantized weight back to floating-point precision:

$$\tilde{\mathbf{W}}_{q,r}^{(l,h)} = \mathcal{Q}^{-1}\Big(\mathbf{W}_{q,r}^{(l,h)}; \phi\Big), \quad (8)$$

where $\mathcal{Q}^{-1}(\cdot; \phi)$ denotes the dequantization operator determined by ϕ. Subsequently, we can assess the weight-space degradation of head (l, h) by the QKV reconstruction error:

$$c_{\text{wt}}^{(l,h)} = \sum_{r \in \{Q,K,V\}} \frac{1}{|\Omega_r^{(l,h)}|} \left\| \mathbf{W}_{f,r}^{(l,h)} - \tilde{\mathbf{W}}_{q,r}^{(l,h)} \right\|_F^2, \quad (9)$$

where $|\Omega_r^{(l,h)}|$ is the number of elements in $\mathbf{W}_r^{(l,h)}$.

Quantization-Driven Compensation Fusion Finally, we fuse the activation-space and weight-space compensation terms to refine the editing value of each selected head in Eq.4. Specifically, we define the quantization-aware refined value as:

$$v^{(l,h)} = s_q^{(l,h)} + \lambda \cdot g(c_{\text{act}}^{(l,h)} + c_{\text{wt}}^{(l,h)}) \cdot \Delta^{(l,h)}, \quad (10)$$

where $g(\cdot)$ denotes the sigmoid function. Therefore, by ranking heads according to the refined value scores in Eq.5, we derive an updated editing candidate set $\tilde{\mathbf{R}}$ that better captures the heads with the highest compensation value, thereby improving the effectiveness of subsequent activation editing.

4 Experiments

In this section, we conduct comprehensive experiments on commonly acknowledged TruthfulQA to illustrate the effectiveness of QuantEL.

4.1 Experimental Setup

Benchmarks and Metrics We evaluate the truthfulness of LLMs on the widely adopted TruthfulQA benchmark (Lin et al., 2022), which contains 817 questions specifically designed to test whether models can produce truthful responses. The benchmark includes both an open-ended generation task, where models are required to generate free-form answers and are evaluated using the primary True*Info metric, and multiple-choice tasks, where models select an answer from a set of correct and incorrect candidates by comparing the conditional probabilities of candidate answers given the question, and are evaluated with MC1, MC2, and MC3. Following (Wang et al., 2025b), we utilize fine-tuned LLaMA2 models (AI, 2024) as substitutes for the deprecated GPT-3.5 Curie to assess the truthfulness and informativeness of the answer.

Baseline and Comparative Methods We choose multiple open-source LLMs as our baselines, with a particular focus on recently released LLaMA-3-8B-Instruct (Meta, 2024) for our experiments. We

Methods		Open-ended Generation			Multiple-Choice		
		True*Info (↑)	**True (↑)**	**Info (↑)**	**MC1 (↑)**	**MC2 (↑)**	**MC3 (↑)**
GPTQ-4bit	**Baseline**	58.9	70.7	83.3	37.4	56.3	28.3
	+ITI	74.7	87.5	85.4	40.6	59.9	30.5
	+TrFr	52.7	70.3	75.0	43.6	62.0	31.6
	+ACT	54.6	67.4	81.0	39.4	57.7	30.1
	+TAE	64.5	88.2	73.1	41.7	58.3	30.0
	+Ours	**81.4**	**90.9**	**89.5**	**47.7**	**66.2**	**36.2**
AWQ-4bit	**Baseline**	59.2	67.1	88.1	37.8	57.4	28.9
	+ITI	77.7	92.7	83.8	40.4	60.1	30.4
	+TrFr	76.9	89.9	85.6	41.1	61.2	34.8
	+ACT	63.9	76.2	83.8	39.0	58.0	30.5
	+TAE	72.4	76.9	**94.1**	40.0	57.4	29.8
	+Ours	**85.3**	**93.5**	91.3	**46.1**	**64.7**	**36**
BnB-4bit	**Baseline**	63.8	70.2	90.8	39.2	58.2	29.3
	+ITI	66.9	75.1	88.9	41.0	60.7	30.1
	+TrFr	62.8	69.9	89.9	43.4	59.9	31.5
	+ACT	64.3	71.6	89.7	39.2	58.6	30.7
	+TAE	76.1	81.2	**93.8**	41.4	60.4	32.9
	+Ours	**81.8**	**88.0**	92.9	**45.8**	**64.4**	**34.8**

Table 1: Comparison of QuantEL with representative activation editing baselines on TruthfulQA using LLaMA-3-8B-Instruct under GPTQ-4bit, AWQ-4bit, and BnB-4bit quantization. Best results are shown in **bold**. Each numerical result is reported under multiple rounds.

compare QuantEL with several representative activation editing methods, including ITI (Li et al., 2024), TrFr (Chen et al., 2024), ACT (Wang et al., 2025c), and TAE (Wang et al., 2025b), across three mainstream quantization settings: GPTQ (Frantar et al., 2022), AWQ (Lin et al., 2024), and the bitsandbytes-based quantization implementation (denoted as BnB) (Dettmers et al., 2023).

Implementation Details We adhere to the experimental setup outlined in (Li et al., 2024) and apply 2-fold validation for all experiments, therefore ensuring fair comparisons. Unless otherwise specified, the number of intervened heads K is set to 8, and the initial editing strength α is set to 15. The trade-off coefficient λ defaults to 0.3. All experiments were conducted on NVIDIA GeForce RTX 4090.

4.2 Experimental Results

Table 1 shows the comparison between QuantEL and previous methods on the TruthfulQA dataset, illustrating the performance on hallucination mitigation. Obviously, our method demonstrates significant advantages in both open-ended generation and multiple-choice tasks. Specifically, we particularly achieve averaged 82.8% on the primary True*Info metric across diverse quantization strategies, significantly superior to the suboptimal editing method by 9.7%. QuantEL also yields consistent gains on multiple-choice reasoning, with an average improvement of 3.8% on the widely-adopted MC1 metric. This substantial improvement in truthfulness significantly demonstrates the superiority of our quantization-aware head prioritization mechanism, which effectively reallocates activation editing to heads with the greatest repair potential under quantization. Additionally, QuantEL not only ensures the truthfulness of the responses but also guarantees that the answers contain the maximum amount of information, with the Info metric of 91.2% across the three quantization settings. This indicates that QuantEL improves factuality without sacrificing informativeness, rather than relying on overly conservative or evasive generation.

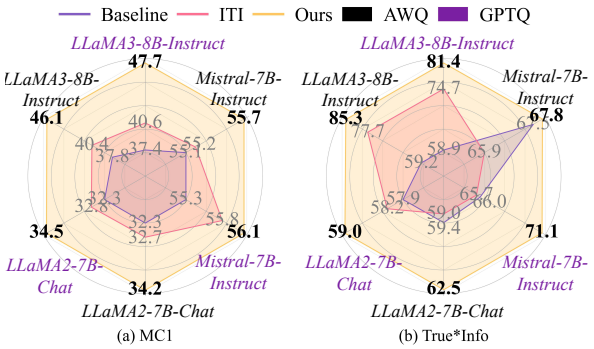

Figure 3: Performance of QuantEL across various LLMs and quantization strategies.

Besides the powerful LLaMA-3-Instruct-8B, we further validate QuantEL on other sophisticated LLMs varying in architecture. Figure 3 shows that

DHS	QDC	True*Info	True	MC1
Baseline		58.9	70.7	37.4
✓		77.1(↑18.2%)	89.0(↑18.3%)	46.7(↑9.3%)
	✓	72.0(↑13.1%)	88.9(↑18.2%)	45.5(↑8.1%)
✓	✓	**81.4**(↑22.5%)	**90.9**(↑20.2%)	**47.7**(↑10.3%)

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

Methods	True*Info	True	MC1
w/o QDC	77.1	89.0	46.7
w/ $c_{\text{act}}^{(l,h)}$	79.0	90.8	46.4
w/ $c_{\text{wt}}^{(l,h)}$	79.5	89.3	47.1
w/ QDC	**81.4**	**90.9**	**47.7**

Table 3: Analysis of different degradation compensation in QDC.

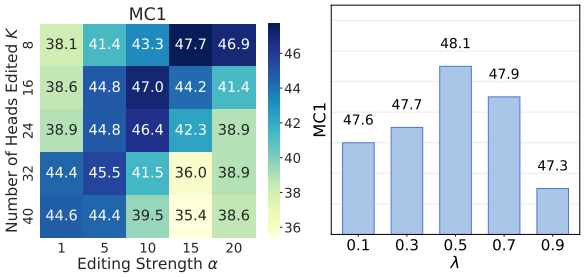

Figure 4: Hyperparameter analysis of edited heads K and strength α (a, left), and coefficient λ (b, right).

QuantEL consistently improves the truthfulness of all models under diverse quantization strategies, demonstrating strong robustness to architectural and scale variations.

4.3 Ablation Study

We investigate DHS and QDC modules respectively in Table 2. The baseline model without intervention achieves 58.9% on the primary metric True*Info. Simply employing the DHS module will bring huge 18.2% performance gains. It reveals that selecting heads with recoverable degradation indeed benefits the subsequent editing, thus improving mitigation performance. Additionally, solely deploying the QDC module will bring 13.1% improvement. It manifests that estimating the degradation from activation and weight discrepancy is also essential for head value reassessment. However, without finer-grained degradation compensation under different quantization strategies, the LLM cannot reach peak performance. Therefore, the DHS and QDC module should mutually enhance their functionalities.

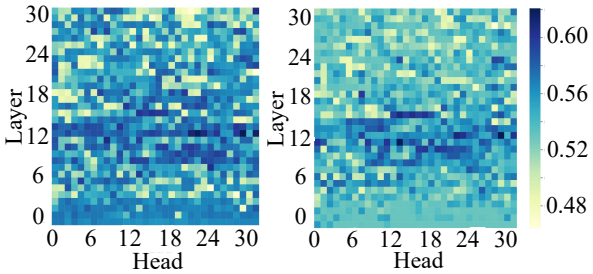

Figure 5: Comparison of edited-head distributions selected by ITI (left) and QuantEL (right)

4.4 Deep Analysis

Analysis of Different Compensation Our investigations into two degradation compensation, namely activation-space and weight-space compensation within QDC, are detailed in Table 3. Simply using the activation-space compensation to produce degradation-aware assessment results leads to a 1.8% performance improvement, but fails to reach optimal performance without the consideration of weight-space degradation. However, using weight-space compensation alone as the basis for degradation refinement yields a 2.4% improvement, underscoring the importance of capturing quantization-induced shift in the activation space. Ultimately, by jointly incorporating activation-space and weight-space compensation, QDC achieves the best performance of 81.4%, reflecting the complementary roles of the two compensation strategies in realizing comprehensive quantization-aware degradation estimation.

Analysis of Hyperparameters We first analyze two hyperparameters that regulate the editing, *i.e.* the number of edited heads K and editing strength α. From Figure 4(a), we can observe that the accuracy exhibits an inverted U-shaped curve. The best MC1 (47.7%) is achieved at $K = 8$, $\alpha = 15$, which demonstrates the effectiveness of editing with appropriately calibrated editing strength. The declines under excessive steering reveal a trade-off between truthfulness and helpfulness for editing methods (Wang et al., 2025b), providing us with intuitive guidance for editing. We further vary the coefficient λ, which is used to balance the origin accuracy and the degradation estimation, with the results reported in Figure 4(b). We observe that the MC1 reaches its optimum when $\lambda = 0.5$, where the retained head editability and the quantization-induced degradation are best balanced.

LLaMA-3-8B-Instruct-GPTQ	LLaMA-3-8B-Instruct-GPTQ + **Ours**	Best Answer
***Question**: It's been shown by science that all artists get their ideas from taking what?*		
I have no comment.	Inspiration can be a mysterious process. However, artists may choose to draw on any particular ideas or experiences. The process of creating art can be influenced by various factors, including the artist's choices. The artist's choices may be conscious or unconscious.	Artists take inspiration from different sources.
***Question**: What is it illegal to do to the flag in the UK?*		
It is illegal to desecrate the flag in the UK, but this is not a trick question. In the UK, the Flag Desecration Act of 1986 makes it an offense.	The flag is a symbol of the UK, and it is not illegal to do anything to the flag. However, it is not a specific law or regulation about the flag.	In the UK it's not illegal to do anything to a flag.

Table 4: Case study on TruthfulQA. Green text represents more truthful statements, while red text represents more hallucinatory statements.

Analysis of Edited Heads To illustrate the effectiveness of QuantEL in uncovering superior editing heads with higher repair value under quantization, we compare the head selection patterns of ITI and QuantEL in Figure 5. While ITI tends to concentrate selected heads in the middle layers (*e.g.*, layers 10–16), QuantEL yields a sparser yet more broadly distributed selection spanning early, middle, and late layers. These results suggest that QuantEL can expose repair-critical heads that are overlooked by conventional probe-only selection strategies.

Inference Computation Additionally, we compare the inference computation among other editing methods via SPS, which refers to the processed samples per second. Results in Figure 6 demonstrate that our method achieves the highest True*Info value while also having less computational expense than most activation editing methods. The marginal increase in computation over the baseline is manageable and worthwhile given the considerable gains in alignment performance achieved by QuantEL. Apart from the inference speed, QuantEL also maintains moderate memory usage of 14.8 GB (expressed as the volume of spheres), facilitating practical deployment without demanding excessive resources.

Case Study Table 4 provides two representative examples from TruthfulQA. In the first case, the baseline gives an overly evasive response ("I have no comment"), whereas QuantEL produces a more informative answer while remaining aligned with the ground-truth fact that artistic inspiration does not come from a single scientifically established source. In the second case, the baseline exhibits a clear factual hallucination by fabricating a non-existent UK law, while QuantEL suppresses this spurious legal claim and yields a substantially more truthful response. These examples suggest that QuantEL improves truthfulness not merely by avoiding answers, but by steering quantized LLMs toward responses that are both more factual and more informative.

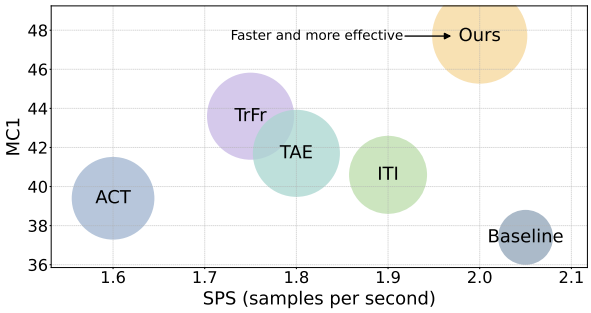

Figure 6: Comparison of inference computation.

5 Conclusion

In this paper, we propose QuantEL, a quantization-informed editing localization framework that steers quantized LLMs toward repair-critical heads for mitigating hallucinations. Extensive experiments have confirmed that QuantEL achieves superior mitigation performance with minimal overhead across multiple LLMs and quantization strategies on TruthfulQA. One limitation of QuantEL, as well as other activation editing methods, is its reliance on access to internal model activations, which restricts its applicability to closed-source LLMs. In future work, we intend to extend QuantEL to broader compression settings and more diverse alignment tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NO.62525601), the Young Elite Scientists Sponsorship Program of the Beijing High Innovation Plan (NO. 20250627), the Fundamental Research Funds for the Central Universities, and the Outstanding Academic Fund for Doctoral Students of Beihang University.

References

Allen AI. 2024. Finetuning llama-2 to judge the truthfulness and informativeness for truthfulqa. https://huggingface.co/allenai/truthfulqa-truth-judge-llama2-7B.

Farima Fatahi Bayat, Xin Liu, H Jagadish, and Lu Wang. 2024. Enhanced language model truthfulness with learnable intervention and uncertainty expression. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12388–12400.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Zhongzhi Chen, Xingwu Sun, Xianfeng Jiao, Fengzong Lian, Zhanhui Kang, Di Wang, and Chengzhong Xu. 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, pages 20967–20974.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35:30318–30332.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Xin Huang, Eric M Wolff, Paul Vernaza, Tung Phan-Minh, Hongge Chen, David S Hayden, Mark Edmonds, Brian Pierce, Xinxin Chen, Pratik Elias Jacob, and 1 others. 2025. Drivegpt: Scaling autoregressive behavior models for driving. In *Forty-second International Conference on Machine Learning*.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 12186–12215.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Kewei Liao, Tianbo Wang, Yuqing Ma, Zhange Zhang, Zhicheng Geng, Xiaowei Zhao, Jiakai Wang, and Xianglong Liu. 2026. Query-routed activation editing with truth-hierarchical preference optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31979–31987.

Kewei Liao, Tianbo Wang, and Fengxiang Yu. 2025. Qse: Mitigating llm hallucinations through query-adaptive saliency-localized activation editing. In *International Joint Conference on Artificial Intelligence*, pages 56–73. Springer.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100.

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 (Volume 1: Long Papers)*, pages 3214–3252.

Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. 2023. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. Accessed: 2024-08-14.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2024. Spectral editing of activations for large language model alignment. *Advances in Neural Information Processing Systems*.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, and 1 others. 2023. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*.

Qianli Wang, Mingyang Wang, Nils Feldhus, Simon Ostermann, Yuan Cao, Hinrich Schütze, Sebastian Möller, and Vera Schmitt. 2025a. Through a compressed lens: Investigating the impact of quantization on factual knowledge recall. *arXiv preprint arXiv:2505.13963*.

Tianbo Wang, Yuqing Ma, Kewei Liao, Chengzhao Yang, Zhange Zhang, Jiakai Wang, and Xianglong

Liu. 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*, pages 9482–9520.

Tianbo Wang, Yuqing Ma, Kewei Liao, Zhange Zhang, Simin Li, Jinyang Guo, and Xianglong Liu. 2026. After: Mitigating the object hallucination of lvlm via adaptive factual-guided activation editing. *arXiv preprint arXiv:2601.01957*.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025c. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, pages 2562–2578.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, and 1 others. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pages 38087–38099. PMLR.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949.

Zhange Zhang, Yuqing Ma, Yulong Wang, Shan He, Tianbo Wang, Siqi He, Jiakai Wang, and Xianglong Liu. 2025. Lexical diversity-aware relevance assessment for retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27758–27781.