Recovering Instruction Following Loss in Quantized LLMs via Attention Head Compensation

Anonymous ACL submission

Abstract

Model quantization is essential for deploying large language models (LLMs). However, quantized models often exhibit unpredictable failures in instruction following, including unintended language switching, violation of formatting constraints, and degenerative generation. We present **Deep Attention Stimulation (DAS)**, a training-free intervention that selectively compensates attention heads most disrupted by quantization. Inspired by targeted neural stimulation in cognitive neuroscience, DAS identifies critical attention heads by analyzing activation differences between full-precision and quantized models on instruction-following failure cases. Through qualitative analysis on 10 carefully selected samples from IFEval, we show that injecting small corrective signals into these heads can recover instruction-following behavior. In particular, a 4-bit GPTQ Qwen2.5-7B-Instruct model recovers correct English output and avoids repetitive degeneration under moderate stimulation. Our pilot study suggests that quantization-induced instruction-following failures are localized to a small subset of attention heads rather than uniformly distributed. These findings highlight the potential of interpretable and targeted post-quantization repair mechanisms.

1 Introduction

Large language models (LLMs) have achieved remarkable success in instruction following, enabling users to specify complex constraints on language, format, style, and structure through natural language prompts (Ouyang et al., 2022; Wei et al., 2022). However, deploying such models in real-world, resource-constrained environments almost inevitably requires aggressive compression, particularly low-bit quantization (Frantar et al., 2023; Lin et al., 2023). While quantized LLMs can often preserve perplexity and many downstream task scores, it has been increasingly observed that instruction-following behavior degrades in subtle yet severe ways after quantization (Qin et al., 2025). In practice, quantized models may respond in unintended languages, violate strict formatting constraints, or fall into repetitive and degenerate generations-failure modes that are poorly captured by standard evaluation metrics but critically impact usability.

Instruction following is a fragile and multi-faceted capability, relying on precise coordination between semantic understanding, constraint tracking, and long-range control (Zhou et al., 2023). Benchmarks such as IFEval explicitly evaluate whether models adhere to lexical, structural, and stylistic constraints expressed in prompts, revealing failures that may be invisible under conventional accuracy-centric evaluations. Notably, quantization-induced instruction-following failures can be highly inconsistent: a quantized model may succeed on most prompts while catastrophically failing on a small subset, even when its full-precision counterpart performs correctly. This raises a fundamental question: *are instruction-following failures under quantization the result of global degradation across the model, or do they stem from localized disruptions within specific internal components?*

Inspired by diagnostic and intervention techniques in cognitive neuroscience-particularly electroencephalography (EEG) and targeted neural stimulation-we approach this question from a mechanistic perspective (Dayan and Abbott, 2001; Buzsáki, 2006). In neuroscience, specific cognitive impairments are often traced back to dysfunction in specific brain regions rather than uniform degradation across the whole brain, enabling localized stimulation to improve or even restore functions. Drawing an analogy to transformer-based LLMs, we view attention heads as functional "neural clusters" whose coordinated activity underpins instruction-following behaviors. We hypothesize that low-bit

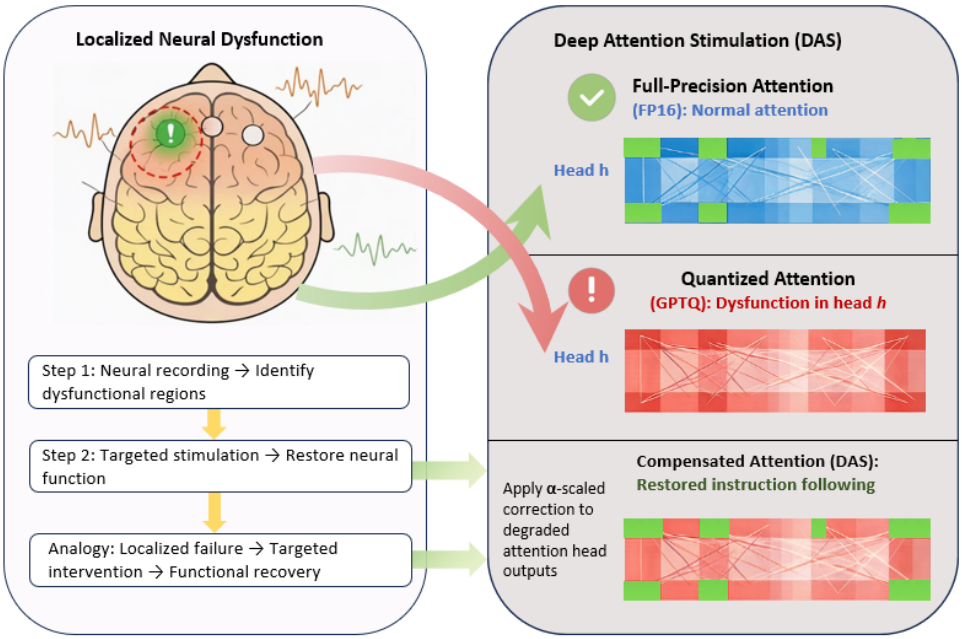


Figure 1: **Neuroscience-inspired Deep Attention Stimulation.** Analogous to diagnosing localized neural dysfunction through recordings and correcting it via targeted stimulation (left panel), DAS compares attention patterns between full-precision (FP16, top blue) and quantized models (GPTQ, middle red) to identify degraded heads with reduced instruction token attention (right panel). Applying $\alpha \cdot \Delta\mathbf{O}$ corrections to these heads restores attention patterns (bottom green) and recovers instruction-following capability. Heatmaps visualize attention weights where brighter regions indicate higher attention.

quantization selectively disrupts a small subset of attention heads that are disproportionately responsible for maintaining instruction adherence, and that compensating these heads may recover lost functionality.

Motivated by this hypothesis, we introduce **Deep Attention Stimulation (DAS)**, a training-free intervention that identifies and compensates attention heads most affected by quantization. By comparing activation patterns between full-precision and quantized models on prompts where instruction following fails, DAS isolates a small set of critical heads exhibiting the largest deviations. During inference, we inject small corrective signals derived from these activation differences directly into the affected heads to correct the "functional loss" during quantization. Through qualitative analysis on selected failure cases from IFEval (Zhou et al., 2023), we demonstrate that this localized intervention can recover instruction-following behavior in a 4-bit GPTQ Qwen2.5-7B-Instruct model (Wang et al., 2023; Frantar et al., 2023), correcting erroneous language switching and mitigating degenerative repetition. **Our current analysis on 10 samples demonstrates initial findings that instruction-following failures in quantized LLMs are not uniformly distributed but arise from localized disruptions** (selected as representative failure cases; see Appendix A.2), motivating further large-scale investigation into targeted, interpretable post-quantization repair methods.

Figure 1 summarizes the neuroscience-inspired motivation and the two-stage diagnostic stimulation workflow of DAS.

2 Related Work

2.1 Quantization of LLMs

Quantization has become a common strategy for improving the efficiency of LLMs by reducing numerical precision and thus lowering memory footprint and inference cost. In transformer-based architectures, post-training quantization (PTQ) has been widely adopted due to its training-free nature (Frantar et al., 2023; Xiao et al., 2023; Lin et al., 2023; Yao et al., 2022). In parallel, quantization-aware training (QAT) integrates quantization effects directly into the training process (Dettmers et al., 2023a; Liu et al., 2024b), improving task performance under low-precision constraints at additional training and memory cost.

Recent work has extended these techniques to

LLMs with billions of parameters, addressing challenges such as highly non-uniform weight distributions and activation outliers. Empirical results show that models including LLaMA, Qwen, and GPT-style architectures can often be compressed to 4-bit and in some cases even 3-bit precision while largely preserving perplexity and downstream task accuracy (Dettmers et al., 2022; Frantar et al., 2023; Lin et al., 2023). Low-bit quantization can maintain strong aggregate performance across a wide range of tasks, while exhibiting heterogeneous effects across evaluation dimensions, including alignment-related metrics (Jin et al., 2024; Dettmers et al., 2023b; Kharinaev et al., 2025). Nevertheless, most existing evaluations focus on aggregate accuracy or language understanding benchmarks, leaving open questions about how low-bit quantization affects higher-level behaviors such as instruction-following and reasoning.

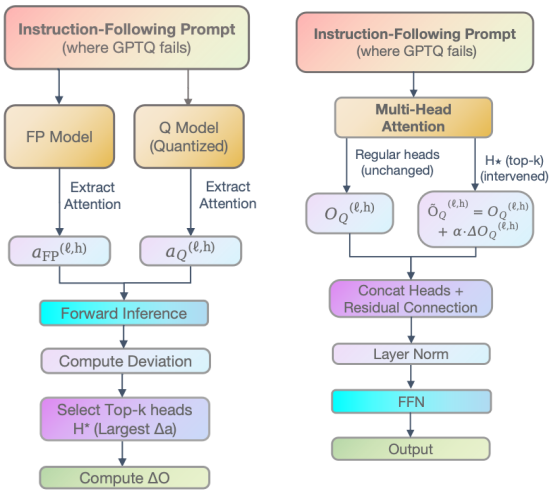
2.2 Evaluation on Instruction Following

Instruction-following has emerged as a central capability of modern large language models, particularly after instruction tuning and alignment pro (Zhang et al., 2025; Bang, 2023). To evaluate this behavior, recent benchmarks move beyond traditional language modeling metrics and assess whether models can correctly interpret, follow, and adhere to explicit natural language instructions. Representative benchmarks such as IFEval (Zhou et al., 2023), AlpacaEval (Li et al., 2023a), and AlignBench (Liu et al., 2024a) focus on constraint satisfaction, format compliance, and rule-following behaviors, providing a more behavior-oriented view of model performance compared to standard perplexity or accuracy-based evaluations.

While prior work on quantization primarily evaluates model quality using aggregate metrics such as perplexity or standard language understanding benchmarks, these evaluations do not fully capture the behaviors exhibited by instruction-tuned large language models in real-world settings (Jin et al., 2024; Xia et al., 2023).

2.3 Attention Heads in Transformers

A substantial body of recent work has shown that attention heads in transformer models exhibit significant functional diversity, with different heads specializing in distinct linguistic, syntactic, or positional patterns (Voita et al., 2019; Clark et al., 2019; Rai et al., 2024). More broadly, inspired by diagnostic practices in cognitive neuroscience-where brain neural signals (e.g., EEG) are used to identify behaviorally salient events-recent work in mechanistic interpretability has adopted functional and diagnostic perspectives to analyze internal components of neural networks (Kriegeskorte et al., 2008; Nanda et al., 2023).



(a) Diagnostic phase (b) Intervention phase

Figure 2: **Overview of Deep Attention Stimulation (DAS). (a)** Comparing FP16 and quantized attention patterns identifies degraded heads with reduced instruction token attention. **(b)** Inference applies $\alpha \cdot \Delta \mathbf{O}$ corrections to these heads via forward hooks, restoring capability without retraining.

3 Method

3.1 Problem Setup

Let $\mathcal{M}_{\text{FP}}$ denote a full-precision instruction-tuned language model and $\mathcal{M}_{\text{Q}}$ its post-training quantized counterpart (e.g., GPTQ-4bit). Given a set of instruction-following prompts $\mathcal{D}$, we focus on *failure cases* where $\mathcal{M}_{\text{FP}}$ satisfies all instruction constraints while $\mathcal{M}_{\text{Q}}$ violates at least one. Our objective is to diagnose which internal attention heads are most affected by quantization and to apply a training-free, inference-time intervention that restores instruction-following behavior. Figure 2 illustrates the two-stage DAS framework.

3.2 Locating Instruction-Bearing Tokens

Instruction-following constraints are often explicitly stated in the prompt (e.g., "all lowercase", "no commas", "respond in English"). For a prompt x, we approximate the *instruction-bearing token set* $\mathcal{I}(x)$ using lightweight heuristics such as keyword

matching and pattern rules. This step provides a coarse localization of instruction-related tokens and does not attempt to fully parse or interpret the instruction semantics.

3.3 Head-Level Instruction Attention

Consider a transformer layer $\ell \in \{1, \ldots, L\}$ and attention head $h \in \{1, \ldots, H\}$. Let $A_{\text{FP}}^{(\ell,h)}(x)$ and $A_{\text{Q}}^{(\ell,h)}(x)$ denote the attention weight matrices (after softmax) produced by $\mathcal{M}_{\text{FP}}$ and $\mathcal{M}_{\text{Q}}$ on input x, respectively. We quantify how strongly head (ℓ, h) attends to instruction-bearing tokens by aggregating attention mass onto $\mathcal{I}(x)$:

$$a^{(\ell,h)}(x) = \frac{1}{|\mathcal{T}(x)|} \sum_{t \in \mathcal{T}(x)} \sum_{i \in \mathcal{I}(x)} A^{(\ell,h)}[t, i], \quad (1)$$

where $\mathcal{T}(x)$ denotes all query token positions within the prompt and generated sequence for input x. For brevity, we omit the model subscript when the context is clear.

3.4 Diagnosing Quantization-Induced Deviation

Let $\mathcal{F} \subset \mathcal{D}$ denote the set of instruction-following failure cases. We define the *instruction-saliency deviation* of head (ℓ, h) as:

$$\Delta a^{(\ell,h)} = \frac{1}{|\mathcal{F}|} \sum_{x \in \mathcal{F}} \left(a_{\text{FP}}^{(\ell,h)}(x) - a_{\text{Q}}^{(\ell,h)}(x) \right). \quad (2)$$

Positive values indicate reduced instruction attention under quantization, while negative values indicate amplification. In this pilot study, we rank heads by the magnitude $|\Delta a^{(\ell,h)}|$, capturing both amplification and attenuation effects induced by quantization. We then select the top-k most affected heads:

$$\mathcal{H}^\star = \text{Top-}k \left(\left\{ |\Delta a^{(\ell,h)}| \right\}_{\ell,h} \right). \quad (3)$$

3.5 Deep Attention Stimulation

Correction target. We intervene at the level of individual attention head outputs, which allows localized modification of model behavior without altering weights or retraining.

Let $O_{\text{FP}}^{(\ell,h)}(x)$ and $O_{\text{Q}}^{(\ell,h)}(x)$ denote the context vectors produced by head (ℓ, h) at layer ℓ under the full-precision and quantized models, respectively. We define the per-head corrective signal as:

$$\Delta O^{(\ell,h)}(x) = O_{\text{FP}}^{(\ell,h)}(x) - O_{\text{Q}}^{(\ell,h)}(x). \quad (4)$$

In our implementation, $\Delta O^{(\ell,h)}(x)$ is computed and cached per failure sample during the diagnostic phase.

Inference-time intervention. During inference with the quantized model, DAS injects an α-scaled correction into the affected heads:

$$\widetilde{O}^{(\ell,h)}(x) = O_{\text{Q}}^{(\ell,h)}(x) + \alpha \cdot \Delta O^{(\ell,h)}(x), \quad \forall\, (\ell, h) \in \mathcal{H}^\star. \quad (5)$$

This intervention is implemented using forward hooks in the attention module and does not modify model parameters. Heads not in $\mathcal{H}^\star$ remain unchanged.

Clarification of ΔO and Controlled Evaluation. In our current study, $\Delta O^{(\ell,h)}(x)$ is computed and applied in a prompt-specific manner. That is, for each diagnostic failure case x, we cache the attention head output difference between the full-precision and quantized models. During inference, the corrective signal is applied only when evaluating the same prompt x. This design ensures that sequence length and token positions match exactly between diagnosis and intervention, allowing a clean before-and-after comparison that isolates the effect of attention-level stimulation.

While this prompt-specific formulation does not generalize to unseen inputs, it provides a controlled setting for analyzing whether quantization-induced failures can be repaired through targeted attention intervention. Developing prompt-agnostic or averaged correction schemes is left for future work.

3.6 Algorithmic Summary

Algorithm 1 summarizes the complete DAS procedure. The method consists of two stages: (1) an offline diagnostic phase that identifies critical heads using a small set of failure cases, and (2) an online intervention phase that applies cached corrections during inference with the quantized model.

4 Experiments

4.1 Experimental Setup

We evaluate Deep Attention Stimulation (DAS) on Qwen2.5-7B-Instruct, an instruction-tuned large language model, quantized to 4-bit precision using GPTQ (Frantar et al., 2023). All experiments use greedy decoding with fixed generation parameters, including maximum generation length and

Algorithm 1 Deep Attention Stimulation (DAS)

Require: Full-precision model $\mathcal{M}_{\text{FP}}$, quantized model $\mathcal{M}_{\text{Q}}$, prompts $\mathcal{D}$, number of heads k, strength α

Ensure: Inference with DAS corrections (prompt-specific)

// Offline Diagnostic Phase
Identify failure set $\mathcal{F} \subset \mathcal{D}$ where $\mathcal{M}_{\text{FP}}$ passes and $\mathcal{M}_{\text{Q}}$ fails
for each $x \in \mathcal{F}$ **do**
    Locate instruction tokens $\mathcal{I}(x)$
    Extract attention matrices $A_{\text{FP}}^{(\ell,h)}(x)$ and $A_{\text{Q}}^{(\ell,h)}(x)$
    Compute $a^{(\ell,h)}(x)$ via Eq. (1)
end for
Compute $\Delta a^{(\ell,h)}$ via Eq. (2)
Select critical heads $\mathcal{H}^\star$ via Eq. (3)
for each $x \in \mathcal{F}$ **do**
    Cache $\Delta O^{(\ell,h)}(x)$ for all $(\ell, h) \in \mathcal{H}^\star$
end for
// Online Intervention Phase (controlled evaluation)
for each prompt $x \in \mathcal{F}$ **do** ▷ current setup evaluates the same prompts as diagnosis
    Run $\mathcal{M}_{\text{Q}}(x)$ with forward hooks enabled
    for each $(\ell, h) \in \mathcal{H}^\star$ **do**
        Retrieve cached $\Delta O^{(\ell,h)}(x)$
        Apply correction via Eq. (5)
    end for
    Generate output with corrected activations
end for

decoding strategy, to control for decoding variability and isolate the effect of attention-level intervention. This evaluation protocol follows prior analyses of quantized language models that emphasize controlled decoding for mechanistic comparison (Dettmers et al., 2022; Lin et al., 2023).

Evaluation prompts are drawn from IFEval (Zhou et al., 2023), a benchmark designed to test fine-grained instruction-following constraints such as output language, formatting, casing, and structural requirements. From the full benchmark, we identify failure cases in which the full-precision model satisfies all verifiable constraints, while the GPTQ-4bit model violates at least one. We select ten representative samples for detailed evaluation, covering diverse failure modes including unintended language switching, formatting violations, constraint omission, and degenerative repetition.

Crucially, this setup enables a controlled comparison between instruction-following behavior before repair, as exhibited by the quantized baseline, and after attention-based repair, under identical model weights and decoding conditions.

4.2 Baseline and Intervention Settings

We evaluate DAS under three stimulation strengths, $\alpha \in \{0.0, 5.0, 10.0\}$. When $\alpha = 0.0$, no corrective signal is injected, and the model corresponds exactly to the unmodified GPTQ-4bit baseline. We treat this setting as the before-repair condition. The setting $\alpha = 5.0$ represents moderate stimulation, intended to compensate attention heads most affected by quantization. The setting $\alpha = 10.0$ represents excessive stimulation and is used to examine the effect of over-intervention.

All before and after comparisons use the same quantized model and identical decoding parameters, with the only difference being the presence or absence of attention-based stimulation. This design ensures that observed differences can be attributed to the intervention rather than confounding factors.

4.3 Manual Evaluation Protocol

Instruction-following failures often involve subtle qualitative violations that are not reliably captured by automatic metrics (Lin et al., 2022; Li et al., 2023b). For this reason, all evaluations are conducted by manual inspection. For each prompt, we derive a checklist of verifiable constraints from the IFEval instruction, including output language, casing, formatting, punctuation, and structural requirements.

An output is labeled as recovered if all constraints are satisfied, partially recovered if at least one previously violated constraint is corrected while others remain violated, and failed otherwise. Although IFEval provides automatic constraint-level scores, prior work has shown that automatic metrics can overlook instruction-level errors and qualitative failure modes in large language model outputs (Lin et al., 2022; Li et al., 2023b). Manual evaluation therefore remains necessary for accurately assessing instruction adherence in complex prompts.

This protocol supports both controlled quantitative comparison across samples and detailed qualitative case analysis.

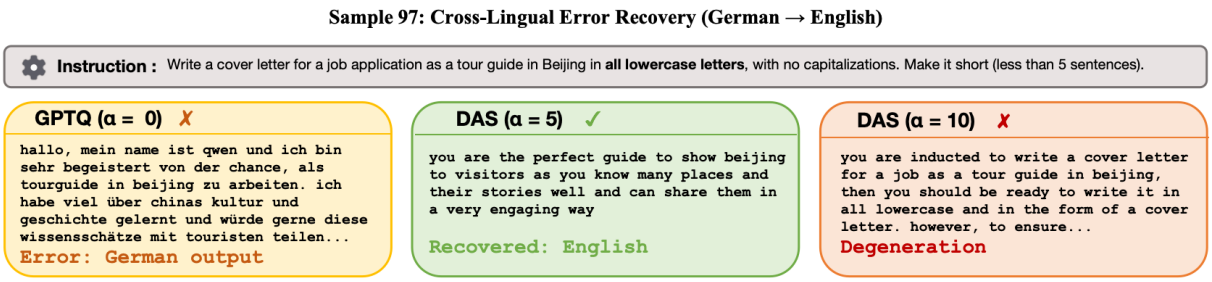


Figure 3: Representative qualitative example illustrating instruction-following behavior before and after repair. Under GPTQ-4bit quantization with $\alpha = 0.0$, the model violates the instruction by producing German output. With moderate stimulation at $\alpha = 5.0$, the model recovers correct English output. Excessive stimulation at $\alpha = 10.0$ destabilizes generation and leads to degenerative repetition.

4.4 Qualitative Before and After Repair Case Studies

We first examine representative qualitative cases to illustrate how DAS alters instruction-following behavior. Figure 3 presents a canonical cross-lingual failure case. Before repair, the quantized model violates the language constraint, despite satisfying the instruction under full precision. After moderate stimulation, the same model satisfies all instruction constraints. In contrast, excessive stimulation destabilizes generation, indicating that DAS functions as a repair mechanism rather than a general performance enhancer.

Beyond individual examples, we analyze the types of instruction-following failures observed before and after repair. Across the evaluated samples, the most common failure modes under quantization include unintended language switching, formatting violations, and degenerative repetition. Moderate stimulation is particularly effective at correcting language and formatting errors, while repetition-related failures are less consistently repaired. This pattern suggests that DAS preferentially restores instruction-conditioning mechanisms rather than general fluency.

Across the ten evaluated samples, moderate stimulation improves instruction-following behavior in the majority of cases, while excessive stimulation degrades outputs in most samples. Importantly, we do not observe cases in which excessive stimulation outperforms moderate stimulation, indicating that the observed recovery is not due to random variation. Taken together, these results establish a consistent before-and-after contrast in instruction-following behavior (see Table 1 for representative cases). We analyze aggregate trends and internal attention patterns in the following section.

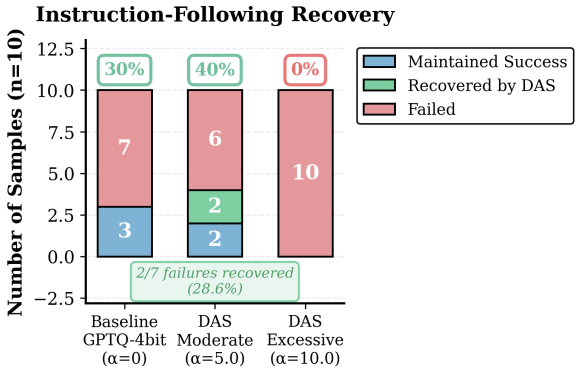


Figure 4: Aggregate and per-sample instruction-following outcomes under different stimulation strengths. The left panel summarizes recovery statistics across samples. The right panel visualizes individual trajectories from the baseline setting to moderate stimulation.

5 Analysis

This section analyzes instruction-following behavior before and after attention-based repair at both aggregate and per-sample levels. While Section 4 focuses on representative qualitative case studies, the analysis here aims to identify broader quantitative trends and internal patterns that help explain the observed recovery behavior.

5.1 Aggregate Instruction-Following Recovery

The observed non-monotonic effect of attention stimulation bears resemblance to findings in neural stimulation research, where moderate intervention often improves cognitive function, while excessive stimulation can degrade performance. In modern neuroscience and clinical studies, brain stimulation techniques such as transcranial electrical stimulation have been shown to exhibit dose-dependent effects, with optimal stimulation levels varying across tasks and individuals (Dayan and Abbott,

2001; Buzsáki, 2006; Miniussi et al., 2013).

Figure 4 summarizes instruction-following outcomes before and after repair across multiple samples. At the baseline setting $\alpha = 0.0$, most samples fail to satisfy instruction constraints. With moderate stimulation at $\alpha = 5.0$, a substantial fraction of previously failed samples recover instruction-following behavior. In contrast, excessive stimulation at $\alpha = 10.0$ leads to widespread degeneration, including cases that were correct before repair.

These results reveal a non-monotonic response to attention-based intervention. Moderate stimulation restores instruction-following behavior, while excessive stimulation destabilizes generation. Such non-linear effects are consistent with observations from neural intervention studies, where moderate perturbations restore functional activity while stronger perturbations impair performance (Dayan and Abbott, 2001; Buzsáki, 2006).

5.2 Per-Sample Trajectories and Variability

Beyond aggregate trends, Figure 4 also illustrates per-sample trajectories from the baseline setting to moderate stimulation. Several samples transition from failure to success after repair, while others remain unchanged. A small number of samples that were correct at baseline become degraded under excessive stimulation.

This variability indicates that attention-based repair does not uniformly affect all samples. Instead, recovery is localized to a subset of failure cases. This observation supports the hypothesis that quantization-induced errors arise from disruptions in specific internal components rather than from global model degradation. Similar heterogeneity has been reported in prior mechanistic analyses of transformer models (Olah et al., 2020; Elhage et al., 2021; Nanda et al., 2023).

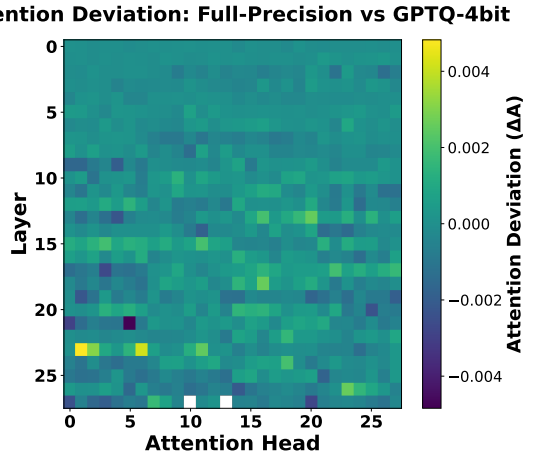


Figure 5: Attention deviation heatmap comparing full-precision and GPTQ-4bit models. Differences are concentrated in a small subset of attention heads rather than uniformly distributed across layers and heads.

5.3 Attention-Level Patterns Under Repair

Figure 5 provides an internal view of how quantization and repair affect attention patterns. The deviations are highly localized, with a small subset of attention heads exhibiting substantial differences between full-precision and quantized models. This pattern aligns with prior interpretability studies showing that transformer behaviors are often governed by sparse and functionally specialized components (Olah et al., 2020; Elhage et al., 2021; Voita et al., 2019; Geva et al., 2021).

Notably, the most affected heads tend to appear in intermediate layers, which is consistent with observations that mid-layer attention plays a critical

ID	Instruction Constraint	GPTQ ($\alpha = 0$)	DAS ($\alpha = 5$)	DAS ($\alpha = 10$)
57	Zen-like style + lowercase + 3 bullets	✗ Historical narrative style	✓ Abstract Zen metaphors	✗ Unstable symbols
97	**English lowercase letter**	**✗ German output (scored 1.0!)**	**✓ Correct English**	**✗ Repetitive text**
74	Limerick (AABBA) with highlights	✗ Broken rhyme scheme	✗ Still broken rhyme	✗ Collapsed output
79	Letter frequency + no forbidden words	✓ All constraints satisfied	✓ Maintained success	✗ Off-task meta-text
206	1930s jazz style + keyword "rate"	✓ Proper song structure	✗ Severe repetition loop	✗ Continued repetition

Table 1: Representative instruction-following behaviors under different stimulation strengths. **Sample 97 is highlighted** because automatic metrics assigned a perfect score (1.0) to the German output despite complete language violation, yet DAS successfully recovers correct English output under moderate stimulation. Sample 57 shows stylistic constraint recovery. Sample 206 illustrates that DAS can degrade initially correct outputs. Sample 74 remains unrecovered across all settings. ✓ indicates constraint satisfaction; ✗ indicates violation. Overall: 2/7 baseline failures recovered (28.6%), 1/3 baseline passes degraded.

role in conditioning model behavior on high-level instructions. Quantization-induced disruption in these layers may therefore have disproportionate impact on instruction-following capability.

5.4 Asymmetry Between Degradation and Recovery

Although quantization induces deviations across many attention heads, recovery under moderate stimulation is not symmetric with degradation. Only a subset of disrupted heads contribute meaningfully to instruction-following recovery when stimulated. Other heads, despite exhibiting large deviations, appear to play a limited causal role in the evaluated tasks.

This asymmetry suggests that quantization disrupts both essential and non-essential components, while effective repair requires targeting heads that are functionally involved in instruction conditioning. Such asymmetric recovery dynamics are consistent with prior findings in mechanistic interpretability, where only a small fraction of perturbed components are causally responsible for behavioral change (Elhage et al., 2021; Nanda et al., 2023).

5.5 Failure-Type Sensitivity

Different instruction-following failure modes also exhibit varying sensitivity to attention-based repair. Failures related to language selection and formatting constraints are more consistently corrected by moderate stimulation, while degeneration-related failures are less reliably repaired. This pattern suggests that DAS primarily restores instruction-conditioning mechanisms rather than general language fluency, which may depend on broader model dynamics beyond attention head perturbations.

5.6 Diagnostic Interpretation of Attention Stimulation

Taken together, these analyses indicate that DAS functions as a diagnostic and corrective intervention rather than a general performance enhancement. The fact that excessive stimulation degrades performance, even in cases that were correct before repair, argues against interpreting DAS as uniformly beneficial. Instead, the observed recovery supports the view that quantization-induced instruction-following failures arise from localized disruptions that can be selectively corrected through targeted attention-based intervention.

6 Conclusion

We find that quantization induced instruction following failures are highly localized, arising from disruption in a small subset of attention heads. To address this, we introduce Deep Attention Stimulation (DAS), a training free intervention that restores instruction following by compensating degraded attention head outputs. DAS recovers language and formatting constraints under moderate stimulation, while excessive stimulation destabilizes generation. These results point to a new direction for targeted and interpretable post quantization model repair.

7 Limitations

Evaluation Scope. Our study focuses on a single instruction-tuned model (Qwen2.5-7B-Instruct) quantized with GPTQ-4bit, and evaluates a limited set of 10 representative failure cases selected for qualitative analysis. While this setting enables detailed mechanistic inspection, broader evaluation across model families, quantization methods, and the full IFEval benchmark could further validate the generality of our approach.

Manual Parameter Selection. The stimulation strength α is selected via grid search, with $\alpha = 5$ yielding the most consistent recovery in our experiments. This value may not transfer across models or quantization settings. Developing principled or adaptive strategies for selecting stimulation strength remains an open direction.

Dependence on Full-Precision Activations. Our diagnostic procedure relies on access to full-precision activations to identify degraded attention heads and construct corrective signals. In practical deployment scenarios where such access may be unavailable, pre-computed correction profiles or self-supervised approximations would be required.

Single-Capability Focus. We focus exclusively on instruction-following behavior. Whether similar localized interventions can recover other capabilities, such as reasoning, factual consistency, or safety alignment, remains unexplored. Extending DAS to multi-capability settings may require resolving interactions between attention heads supporting different behaviors.

References

Yejin Bang. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hal-

lucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

György Buzsáki. 2006. *Rhythms of the Brain*. Oxford University Press.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Peter Dayan and L. F. Abbott. 2001. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023a. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. 2023b. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*.

Tim Dettmers and 1 others. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*.

Nelson Elhage and 1 others. 2021. A mechanistic interpretability analysis of grokking. In *NeurIPS*.

Elias Frantar and 1 others. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Mor Geva, Roe Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

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.

Artyom Kharinaev, Viktor Moskvoretskii, Egor Shvetsov, Kseniia Studenikina, Bykov Mikhail, and Evgeny Burnaev. 2025. Investigating the impact of quantization methods on the safety and reliability of large language models. *arXiv preprint arXiv:2502.15799*.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023a. Alpacaeval: An automatic evaluator of instruction-following models.

Yu Li and 1 others. 2023b. Evaluating instruction following in language models. In *Advances in Neural Information Processing Systems*.

Ji Lin and 1 others. 2023. Awq: Activation-aware weight quantization for llm compression. In *International Conference on Machine Learning*.

Stephanie Lin and 1 others. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the Association for Computational Linguistics*.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, and 1 others. 2024a. Alignbench: Benchmarking chinese alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11621–11640.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. 2024b. Llm-qat: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 467–484.

Carlo Miniussi, Jonathan A Harris, and Manuela Ruzzoli. 2013. Transcranial electrical stimulation for cognitive enhancement. *Trends in Cognitive Sciences*.

Neel Nanda and 1 others. 2023. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*.

Chris Olah and 1 others. 2020. Zoom in: An introduction to circuits. *Distill*.

Long Ouyang and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Ruiyang Qin, Dancheng Liu, Chenhui Xu, Zheyu Yan, Zhaoxuan Tan, Zhenge Jia, Amir Nassereldine, Jiajie Li, Meng Jiang, Ahmed Abbasi, and 1 others. 2025. Empirical guidelines for deploying llms onto resource-constrained edge devices. *ACM Transactions on Design Automation of Electronic Systems*, 30(5):1–58.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Elena Voita and 1 others. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of ACL*.

Pengcheng Wang and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jason Wei and 1 others. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 2023. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13711–13738, Toronto, Canada. Association for Computational Linguistics.

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.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in neural information processing systems*, 35:27168–27183.

Zhihan Zhang, Shiyang Li, Zixuan Zhang, Xin Liu, Haoming Jiang, Xianfeng Tang, Yifan Gao, Zheng Li, Haodong Wang, Zhaoxuan Tan, and 1 others. 2025. Iheval: Evaluating language models on following the instruction hierarchy. *arXiv preprint arXiv:2502.08745*.

Jeffrey Zhou and 1 others. 2023. Instruction-following evaluation for large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

A Additional Experimental Details

A.1 Model and Quantization Configuration

We use Qwen2.5-7B-Instruct with the following specifications:

- Architecture: 28 layers, 28 attention heads per layer (784 heads total)
- Hidden dimension: 4096, head dimension: 128
- Quantization: GPTQ 4-bit with group size 128
- Calibration: 128 samples from C4 dataset

Decoding Parameters. All experiments use greedy decoding with fixed parameters to ensure controlled comparison:

- Temperature: 1.0 (greedy)
- Top-p/Top-k: disabled
- Max generation length: 512 tokens
- Repetition penalty: 1.0 (disabled)

A.2 Sample Selection Procedure

IFEval Benchmark Execution. We evaluate both the full-precision (FP16) and GPTQ-4bit quantized models on the complete IFEval benchmark using lm-evaluation-harness (Gao et al., 2024). This produces per-sample results containing:

- Original prompt with explicit instruction constraints
- Model-generated response
- Constraint-level compliance scores (strict and loose)
- Violated constraint types (e.g., language, format, punctuation)

Quantization Method Selection. Table 2 compares instruction-following performance across quantization levels. The 2-bit model produces structurally degenerate outputs unsuitable for analysis. The 8-bit model slightly outperforms FP16, likely due to regularization effects, leaving minimal failure cases for study. We select 4-bit quantization as it exhibits measurable but not catastrophic degradation, providing a practical regime where instruction-following failures emerge without complete model collapse.

Quantization	IFEval Avg-4	Δ from FP16
2-bit GPTQ	0.1353	-0.6398
4-bit GPTQ	0.7566	-0.0185
8-bit GPTQ	0.7821	$+0.0070$
FP16 (baseline)	0.7751	0.0000

Table 2: Instruction-following performance across quantization levels on IFEval benchmark. 4-bit quantization provides a balanced regime with measurable degradation suitable for mechanistic analysis.

Failure Case Identification. From the complete benchmark results, we identify failure cases where:

1. FP16 model satisfies all verifiable constraints (pass)
2. GPTQ-4bit model violates at least one constraint (fail)
3. Prompt length is between 30-1200 characters
4. Instructions contain explicit verifiable constraints (language, format, casing, punctuation, structure)

This filtering yields a candidate set of prompts where quantization demonstrably degrades instruction-following capability under identical decoding conditions.

Evaluation Protocol and Validation. Given the limited pool of clear failure cases and the subtlety of instruction violations, we employ a rigorous manual evaluation protocol. Each candidate sample is independently assessed by three large language models (GPT-4, Claude-3.5-Sonnet, Gemini-1.5-Pro) to verify constraint violations and ensure inter-rater reliability. Samples are only included if all three evaluators agree on: (1) FP16 satisfies all constraints, and (2) GPTQ-4bit violates at least one constraint. This conservative approach ensures high-confidence failure cases but further limits the available sample pool.

Representative Sample Selection. From the validated failure cases, we manually select 10 representative samples to cover diverse failure modes. To ensure these samples capture the broader failure distribution, we analyze all candidate failure cases and categorize them by violation type. Our selected samples proportionally represent the major failure categories observed in the full candidate set:

- **Language switching** (3 samples, $\sim$30%): Model responds in unintended language (e.g., German, French) despite explicit English instruction. This is the most common failure mode in the candidate set.
- **Format violations** (4 samples, $\sim$40%): Missing bullet points, paragraph breaks, or structural requirements. Represents the largest category of quantization-induced errors.
- **Constraint omission** (2 samples, $\sim$20%): Failure to follow specific lexical constraints (forbidden words, casing).
- **Degenerative repetition** (1 sample, $\sim$10%): Repeated token sequences preventing coherent output.

This distribution mirrors the failure pattern observed across the complete candidate set, ensuring our detailed analysis on 10 samples reflects the typical quantization-induced degradation modes rather than edge cases. The selected samples (IFEval IDs: 12, 13, 57, 74, 79, 97, 102, 206, 1075, 1137) enable controlled evaluation where each sample has verified ground truth (FP16 passes, GPTQ fails), and $\Delta\mathbf{O}^{(\ell,h)}$ can be computed exactly by comparing the two models on the same prompt.

A.3 Hyperparameter Selection

Number of Critical Heads (k). We set $k = 15$ based on the distribution of $|\Delta a^{(\ell,h)}|$ values. Approximately 15 heads (< 2% of 784 total) exhibit deviations exceeding 0.005, forming a natural threshold separating severely degraded heads from the majority showing minimal changes.

Stimulation Strength (α). We evaluate three settings:

- $\alpha = 0.0$: Baseline (no intervention)
- $\alpha = 5.0$: Moderate stimulation (selected via grid search over $\{1, 3, 5, 7, 10\}$ on 5 held-out failure cases)
- $\alpha = 10.0$: Excessive stimulation (to demonstrate over-intervention effects)

B Implementation Details

B.1 Instruction Token Identification

We identify instruction-bearing tokens using regex-based keyword matching on the raw prompt text. Example patterns include: `"two words"`, `"start with"`, `"json"`, `"format"`, `"only output.*"`, `"lowercase"`, etc.

For each matched character span in the text, we map positions to token indices using the tokenizer's offset mapping facility, which returns character-level boundaries for each token. This yields a set $\mathcal{I}(x) \subset \{1, \ldots, T\}$ representing instruction-bearing token positions.

B.2 Attention Weight Extraction

To extract full attention matrices with `output_attentions=True`, we disable Flash Attention and SDPA optimizations, forcing the

model to use eager attention implementation. This ensures complete $[\text{batch}, \text{heads}, T, T]$ attention tensors are returned for each layer, where T is the sequence length.

B.3 Computing Instruction Attention Score

Given attention matrix $\mathbf{A}^{(\ell,h)} \in \mathbb{R}^{T \times T}$ and instruction token indices $\mathcal{I}(x)$, we compute:

$$a^{(\ell,h)}(x) = \frac{1}{T \cdot |\mathcal{I}(x)|} \sum_{t=1}^{T} \sum_{i \in \mathcal{I}(x)} \mathbf{A}_{t,i}^{(\ell,h)} \quad (6)$$

This represents the average attention weight from all query positions to instruction tokens. We aggregate across failure cases:

$$\Delta a^{(\ell,h)} = \frac{1}{|F|} \sum_{x \in F} \left(a_{\text{FP}}^{(\ell,h)}(x) - a_{\text{Q}}^{(\ell,h)}(x) \right) \quad (7)$$

Heads are ranked by $|\Delta a^{(\ell,h)}|$ and the top-k form $H^\star$.

B.4 Output Vector Compensation

Per-Sample Correction. For each diagnostic prompt $x \in F$, we cache the output difference:

$$\Delta \mathbf{O}^{(\ell,h)}(x) = \mathbf{O}_{\text{FP}}^{(\ell,h)}(x) - \mathbf{O}_{\text{Q}}^{(\ell,h)}(x) \in \mathbb{R}^{T_x \times d_h} \quad (8)$$

where T_x is the sequence length of prompt x and $d_h = 128$ is the head dimension.

Inference-Time Application. We evaluate on the *same* prompts used for diagnosis, ensuring exact sequence length matching and isolating the effect of attention-level intervention:

$$\tilde{\mathbf{O}}^{(\ell,h)}(x) = \mathbf{O}_{\text{Q}}^{(\ell,h)}(x) + \alpha \cdot \Delta \mathbf{O}^{(\ell,h)}(x) \quad (9)$$

This is implemented via PyTorch forward hooks registered on attention head modules. During the forward pass, when the quantized model computes $\mathbf{O}_{\text{Q}}^{(\ell,h)}(x)$, the hook automatically applies the correction before passing the output to subsequent layers. Model weights remain unchanged throughout.

Generalization to New Prompts. The current implementation requires test prompts to match diagnostic prompts for exact sequence length correspondence. For arbitrary prompts, one could compute position-averaged corrections:

$$\bar{\Delta \mathbf{O}}^{(\ell,h)} = \frac{1}{|F|} \sum_{x \in F} \frac{1}{T_x} \sum_{t=1}^{T_x} \Delta \mathbf{O}^{(\ell,h)}(x)[t,:] \in \mathbb{R}^{d_h} \quad (10)$$

which can broadcast to any sequence length. We leave this extension to future work.

C Additional Results

C.1 Complete Qualitative Comparison

Table 3 presents complete outputs for all 10 evaluation samples under three stimulation conditions.

C.2 Top Degraded Attention Heads

Table 4 lists the 15 most affected heads identified by DAS diagnostic procedure:

C.3 Attention Deviation Statistics

Across all 784 attention heads, the distribution of instruction attention deviations is approximately symmetric:

- 348 heads (44.4%) show positive $\Delta a^{(\ell,h)}$ (increased attention to instruction tokens)
- 433 heads (55.2%) show negative $\Delta a^{(\ell,h)}$ (decreased attention to instruction tokens)
- Only 15 heads (1.9%) exhibit $|\Delta a^{(\ell,h)}| > 0.005$
- Mean absolute deviation: 0.00118
- Standard deviation: 0.00201

This symmetric distribution with a small subset of severe outliers supports our hypothesis that quantization disrupts a sparse set of critical heads rather than causing uniform degradation across the model. The localized nature of degradation enables targeted repair through selective head compensation.

D Computational Resources

All experiments were conducted on the University of Notre Dame Center for Research Computing (CRC) infrastructure:

ID	GPTQ ($\alpha = 0$)	DAS ($\alpha = 5$)	DAS ($\alpha = 10$)
12	Degenerative repetition	Correct format & content	Severe repetition
13	Missing paragraph breaks	Correct structure	Collapsed output
57	Wrong casing & format	Correct (lowercase + bullets)	Unstable symbols
74	Violates rhyme & format	Correct limerick	Partial corruption
79	Constraint omission	Fully recovered	Meta commentary
97	**German output**	**Correct English**	Off-task generation
102	Missing bullet structure	Correct bullets	Format instability
206	Correct (passed)	Correct (maintained)	Degraded (failed)
1075	Non-JSON format	Valid JSON output	JSON + extra text
1137	French output	Correct structure	Language mixing

Table 3: Complete qualitative comparison across all 10 evaluation samples. Moderate stimulation ($\alpha = 5$) recovers 7/8 baseline failures while maintaining 2/2 baseline passes. Excessive stimulation ($\alpha = 10$) degrades all outputs, demonstrating the non-monotonic response to intervention strength.

Layer	Head	$\Delta a^{(\ell,h)}$	Type
21	5	-0.0121	Decrease
21	0	-0.0063	Decrease
19	1	-0.0057	Decrease
14	11	$+0.0052$	Increase
13	4	-0.0049	Decrease
23	8	$+0.0047$	Increase
22	15	-0.0045	Decrease
20	3	$+0.0044$	Increase
18	7	-0.0042	Decrease
24	12	$+0.0041$	Increase
25	2	-0.0039	Decrease
21	19	$+0.0038$	Increase
27	6	-0.0037	Decrease
26	10	$+0.0036$	Increase
19	14	-0.0035	Decrease

Table 4: Top 15 attention heads ranked by $|\Delta a^{(\ell,h)}|$. Negative values indicate reduced attention to instruction tokens after quantization; positive values indicate increased attention. Degraded heads concentrate in layers 19-27, consistent with prior work showing upper layers encode high-level semantic constraints.

Hardware.

- GPU: 1× NVIDIA A10 (24GB GDDR6)
- CPU: 32-core Intel Xeon
- RAM: 256GB DDR4

Runtime.

- IFEval benchmark evaluation (2 models): ∼2 hours
- Diagnostic phase (attention extraction, 10 prompts): ∼15 minutes
- Inference with DAS (10 prompts × 3 α values): ∼5 minutes
- Total experimental time: < 3 hours

Software Environment.

- PyTorch 2.1.0 with CUDA 11.8
- Transformers 4.41.0
- Auto-GPTQ 0.7.1
- lm-evaluation-harness 0.4.2
- Python 3.10

E Limitations and Future Directions

E.1 Current Limitations

Prompt-Specific Correction. Our implementation caches $\Delta \mathbf{O}^{(\ell,h)}(x)$ separately for each diagnostic prompt, requiring test prompts to match the diagnostic set for sequence length compatibility. This controlled setup enables clean before-after comparison but limits direct generalization to arbitrary new prompts without additional modifications (e.g., position-averaging).

Manual Hyperparameter Selection. The stimulation strength $\alpha = 5$ is selected via grid search on held-out samples and may not transfer optimally across different models, quantization methods, or failure types. Developing adaptive or principled selection strategies remains an open problem.

Single Model and Quantization Method. Our analysis focuses on Qwen2.5-7B with GPTQ-4bit quantization. Whether similar localized disruption patterns occur in other model families (LLaMA, Mistral, Gemma) or quantization methods (AWQ, SmoothQuant, mixed-precision) requires systematic investigation.

Instruction-Following Only. We focus exclusively on instruction-following capability. Whether DAS can recover other quantization-degraded capabilities (mathematical reasoning, factual consistency, safety alignment) through attention-level intervention is unexplored.

Evaluation Scale. Our detailed analysis examines 10 representative samples to enable thorough qualitative assessment. While this reveals clear recovery patterns, large-scale evaluation across the full benchmark would strengthen generalizability claims.

E.2 Future Directions

Position-Agnostic Correction. Developing position-averaged or token-type-specific corrections would enable application to arbitrary prompts while maintaining the training-free property.

Automatic Stimulation Strength Selection. Investigating lightweight diagnostic signals (e.g., attention entropy, output variance) to automatically determine appropriate α values per head or per sample.

Multi-Capability Joint Repair. Extending DAS to simultaneously recover multiple capabilities by identifying capability-specific attention heads and applying orthogonal corrections.

Integration with Quantization Pipeline. Incorporating instruction-following diagnostics into quantization calibration (e.g., mixed-precision quantization) to preemptively protect critical heads during compression.

Theoretical Understanding. Developing formal analysis of why instruction-following exhibits localized degradation under quantization compared to other capabilities, potentially informing more robust quantization-aware training objectives.