From Signal Degradation to Computation Collapse: Uncovering the Two Failure Modes of LLM Quantization

Chenxi Zhou[1,2], Pengfei Cao[2,3]*, Jiang Li[4], Bohan Yu[1,2], Jinyu Ye[2], Jun Zhao[2,3], Kang Liu[2,3]*

[1]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
[2]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
[4]College of Computer Science, Inner Mongolia University
zhouchenxi2025@ia.ac.cn, {pengfei.cao, jzhao, kliu}@nlpr.ia.ac.cn

Abstract

Post-Training Quantization (PTQ) is critical for the efficient deployment of Large Language Models (LLMs). While 4-bit quantization is widely regarded as an optimal trade-off, reducing the precision to 2-bit usually triggers a catastrophic "performance cliff." It remains unclear whether the underlying mechanisms differ fundamentally. Consequently, we conduct a systematic mechanistic analysis, revealing two qualitatively distinct failure modes: Signal Degradation, where the computational patterns remain intact but information precision is impaired by cumulative error; and Computation Collapse, where key components fail to function, preventing correct information processing and destroying the signal in the early layers. Guided by this diagnosis, we conduct mechanism-aware interventions, demonstrating that targeted, training-free repair can mitigate Signal Degradation, but remains ineffective for Computation Collapse. Our findings provide a systematic diagnostic framework for PTQ failures and suggest that addressing Computation Collapse requires structural reconstruction rather than mere compensation.

1 Introduction

Post-Training Quantization (PTQ) has emerged as a crucial technique for efficient Large Language Model (LLM) deployment. In practice, 4-bit quantization is often regarded as an optimal trade-off (Jin et al., 2024), achieving significant compression with acceptable performance loss. However, reducing the precision to 2-bit with common methods (e.g., GPTQ (Frantar et al., 2023)) usually triggers a catastrophic "performance cliff," particularly in tasks requiring precise factual knowledge. Since factual recall forms the foundation of LLM capabilities, this collapse signals a fundamental breakdown that requires deep investigation.

Existing research on PTQ spans three primary directions. **The first focuses on macroscopic evaluation**, measuring how much performance drops on diverse downstream tasks (Li et al., 2024; Jin et al., 2024; Liu et al., 2025a). **The second direction pursues algorithmic refinement**, employing numerical optimization strategies such as outlier suppression (Lin et al., 2024) or rotation matrices (Tseng et al., 2024) to reduce errors. However, these two directions share a common limitation. They primarily focus on quantifying the performance degradation or minimizing numerical error, but overlook why the model's internal mechanism fails. They treat the quantization damage as a numerical issue rather than investigating the disruption of knowledge storage and recall.

The third stream involves preliminary mechanistic exploration. Common approaches identify critical modules by analyzing layer or component sensitivity (Namburi et al., 2023; Zhang et al., 2025a; Xiao et al., 2025; Dumitru et al., 2025), while deeper studies attribute failures to the "RMSNorm Reversal" effect (Chang et al., 2025). However, these insights remain fragmented, lacking a systematic mechanistic interpretation of the failure modes. Despite these efforts, we still cannot explain why the "performance cliff" exists: *Is the catastrophic failure under common 2-bit merely a quantitative aggravation of 4-bit degradation, or does it mark a qualitative shift to a fundamentally distinct mechanism?*

To answer this, we conduct an in-depth mechanistic analysis. We first trace the layer-wise information flow and causal pathways to investigate whether the knowledge signal exists and propagates correctly. Based on these observations, we reveal two qualitatively distinct PTQ failures. Using standard PTQ settings as representative cases, we propose the **Two Failure Modes Hypothesis**:

*Corresponding authors

Findings of the Association for Computational Linguistics: ACL 2026, pages 23204–23222

- **Failure Mode I: Signal Degradation.** The model's computational patterns remain largely intact. Quantization error acts as cumulative noise that impairs information precision.
- **Failure Mode II: Computation Collapse.** The quantization error is severe enough to fundamentally damage the functionality of key components. Information cannot be processed correctly and is completely destroyed in the early layers.

We validate this hypothesis through a systematic analysis. We examine the functionality of critical components and analyze the internal structure of the representation space. This analysis confirms that Signal Degradation involves functional but impaired components, whereas Computation Collapse stems from a fundamental structural breakdown.

Finally, guided by the diagnosis, we design targeted intervention experiments. We demonstrate that Signal Degradation can be repaired by targeted, training-free strategies. In contrast, Computation Collapse is systemic, where even advanced low-rank compensation remains ineffective, necessitating structural reconstruction (e.g., fine-tuning).

Overall, the main contributions of this work can be summarized as follows:

- We propose a systematic interpretability analysis framework, providing a general approach to diagnose performance decline under quantization.
- We identify two distinct failure modes, Signal Degradation and Computation Collapse, demonstrating that they differ qualitatively rather than merely in severity.
- We clarify the optimization strategies for different failure modes, suggesting that while degradation benefits from targeted repair, collapse requires structural reconstruction rather than mere compensation.

2 Related Work

2.1 Post-Training Quantization

Post-Training Quantization (PTQ) compresses LLMs efficiently, but the primary challenge lies in handling activation outliers. To mitigate this, methods have evolved from simple rounding to sophisticated numerical transformations. Early weight-only methods like GPTQ (Frantar et al., 2023) minimize reconstruction error using Hessian information. Techniques like AWQ (Lin et al., 2024) and SmoothQuant (Xiao et al., 2023) perform channel-wise scaling to suppress outliers, while recent approaches such as QuIP# (Tseng et al., 2024) and SpinQuant (Liu et al., 2025b) employ rotation matrices to flatten activation distributions.

Despite their success in reducing statistical errors like MSE, these methods remain limited to a numerical perspective. By focusing strictly on aligning the output distribution with the full-precision baseline, they overlook internal behaviors and fail to explain how the underlying computational mechanisms change under quantization.

2.2 Mechanistic Analysis of Quantization

Mechanistic interpretability offers tools to reverse-engineer model behaviors, such as decoding hidden states via Logit Lens (nostalgebraist, 2020) or locating knowledge via Causal Tracing (Meng et al., 2022). However, the application of these powerful diagnostic tools to investigate the internal mechanics of quantized models remains preliminary.

Prior work in quantization analysis has largely focused on component sensitivity, identifying fragile layers or modules based on Hessian spectra or weight magnitudes (Zhang et al., 2025a; Dong et al., 2020). More recently, researchers have extended mechanistic analysis to specific model capabilities, such as analyzing the compromise of refusal mechanisms (Chhabra and Khalili, 2025), shifts in truthfulness (Fu et al., 2025), or the unintended recovery of unlearned knowledge (Zhang et al., 2025b). However, these studies remain fragmented, focusing on isolated tasks or behaviors. Our work aims to provide a systematic mechanistic explanation for quantization failures.

3 Two Failure Modes Hypothesis

3.1 Experimental Setup

Models and Quantization. We conduct our primary analysis on Llama-3.1-8B (Grattafiori et al., 2024). To ensure generalizability, we validate findings on Qwen3-8B (Yang et al., 2025), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Gemma-2-9B-it (Team et al., 2024). We select GPTQ (Frantar et al., 2023) as the primary baseline as it is the most widely adopted weight-only PTQ method. We contrast 4-bit (the PTQ sweet-spot) and 2-bit (typically unusable) to investigate their fundamentally distinct degradation behaviors, providing 8-bit and 3-bit results for context. Algorithmic generalizability is further validated using AWQ in Appendix D.

Datasets and Task. We evaluate factual knowledge recall using Pararel (Elazar et al., 2021) (39 relation types). It is deliberately selected because factual recall is a foundational capability, and its strict `<subject>-<relation>-<target>` structure provides fixed token positions, facilitating precise mechanistic diagnosis. Relations are mapped to standardized templates for next-token prediction (Appendix A.2). Generalizability to broader tasks (MMLU, GSM8K) is verified in Appendix E.

Analysis Subsets. To specifically investigate quantization-induced failures, we partition the dataset for each model based on FP16 and 4-bit performance into two core subsets: the Robust Subset (`fp_and_4bit_correct`) and the Failure Subset (`fp_correct_4bit_wrong`). We do not partition for 2-bit models as they universally fail. All subsequent mechanistic comparisons are performed on these subsets.

3.2 Phenomenological Evidence

Performance Cliff. We conduct a multi-prompt robustness evaluation (see Appendix A.2) on the factual recall task. Figure 1 illustrates a pronounced "performance cliff". The degradation from FP16 to 4-bit is gradual, maintaining usability. Conversely, the transition to 2-bit triggers a catastrophic collapse where accuracy plummets to zero. This sharp discontinuity suggests that 2-bit quantization represents a distinct failure state rather than a mere lower-precision version of 4-bit.

Rank Drop vs. Collapse. To analyze the nature of these errors, we examine the rank of the correct answer in the final output distribution (Figure 2). 4-bit primarily leads to an "answer rank drop", where the correct answer shifts downward but typically remains within the top tier (e.g., Top-5). This indicates that the model retains the correct information despite reduced confidence. In contrast, 2-bit results in an "answer rank collapse". The rank falls to thousands, almost random guessing. Qualitatively, 2-bit models collapse into generating high-frequency stop words (e.g., "the", "."), reflecting a complete failure in knowledge recall.

3.3 Layer-wise Knowledge Probing

To investigate the internal status underlying these macroscopic differences, we examine whether a decodable knowledge signal exists within the intermediate states. We employ the logit lens (nostalgebraist, 2020) to project the hidden state $h^{(l)}$ at layer l directly into the vocabulary space via the unembedding matrix W_U. Figure 3 traces the layer-wise change of the correct token's probability and rank, revealing distinct dynamics.

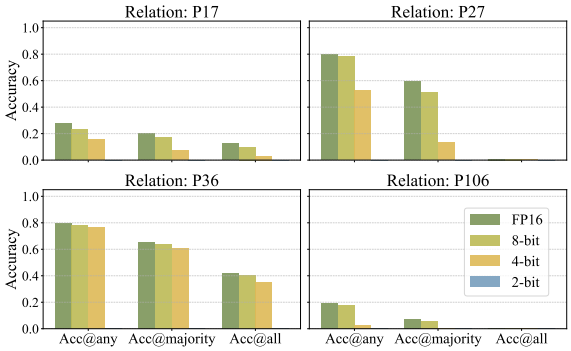


Figure 1: Multi-prompt factual recall accuracy of Llama3.1-8B under different quantization levels on four Pararel relations. We report Accuracy@any ($\geq$1 correct), @majority (>50%), and @all (100%).

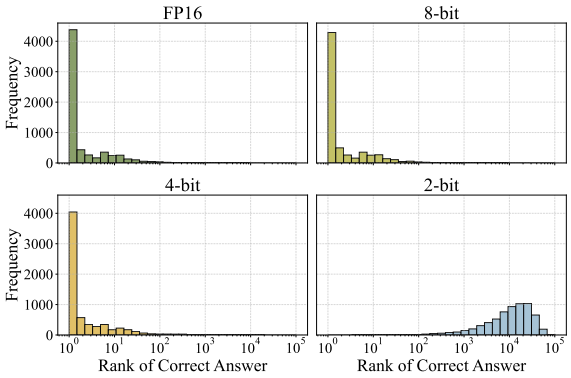


Figure 2: Distribution of the rank of the correct answer for Llama3.1-8B under different quantization levels on four Pararel relations (P17, P27, P36, P106).

Signal Absence. The 2-bit models exhibit a consistent failure to form an effective knowledge signal. As shown in the red curves in Figure 3, the probability of the correct answer remains near zero throughout all layers, and its rank stays extremely low (in the tens of thousands). This indicates that the knowledge signal is never successfully generated during the computation process.

Signal Degradation. In contrast, 4-bit models demonstrate an observable knowledge signal. In the Robust Subset (Fig. 3a, c), the signal closely tracks the FP16 baseline. Even in the Failure Subset (Fig. 3b, d) where the model ultimately fails, the signal still emerges in mid-to-late layers but with reduced intensity. The probability curve shows lower confidence, and the rank improves more slowly than in FP16. This characterizes 4-bit failure as signal degradation, where the correct signal is present

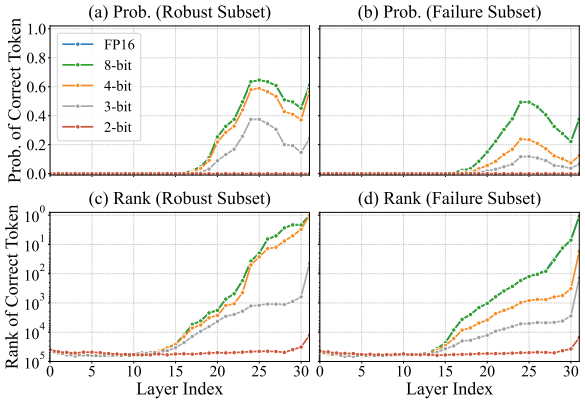
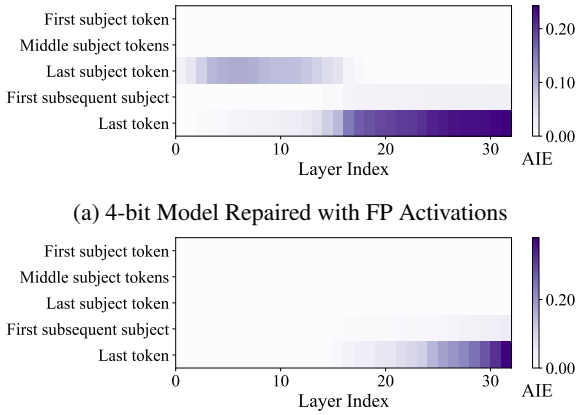


Figure 3: Layer-wise change of probability and rank. FP16 nearly overlaps with 8-bit.



(a) 4-bit Model Repaired with FP Activations

(b) 2-bit Model Repaired with FP Activations

Figure 4: Cross-model activation repair on the Failure Subset. The heatmap values represent the Average Indirect Effect (AIE), defined as the increase in the correct token's prediction probability.

but ultimately overtaken by the noise, unlike the complete absence seen in 2-bit models.

3.4 Causal Analysis of Information Flow

While Section 3.3 analyzes the existence of knowledge signals, it remains unclear whether the causal mechanism for processing them is intact. To distinguish whether the information flow is merely impaired or fundamentally broken, we employ causal activation patching (Heimersheim and Nanda, 2024) to assess the integrity of the information pathway.

(1) Cross-Model Repair (Sufficiency). We adapt Causal Tracing (Meng et al., 2022) to test signal sufficiency. We replace the residual stream state $h_Q^{(l,t)}$ (i.e., the layer output at layer l, token position t) in the quantized model with the corresponding "clean" activation $h_{FP}^{(l,t)}$ from the FP16 model. If this injection restores the correct prediction, it proves that the injected signal is sufficient to restore the output and the downstream pathway remains functional.

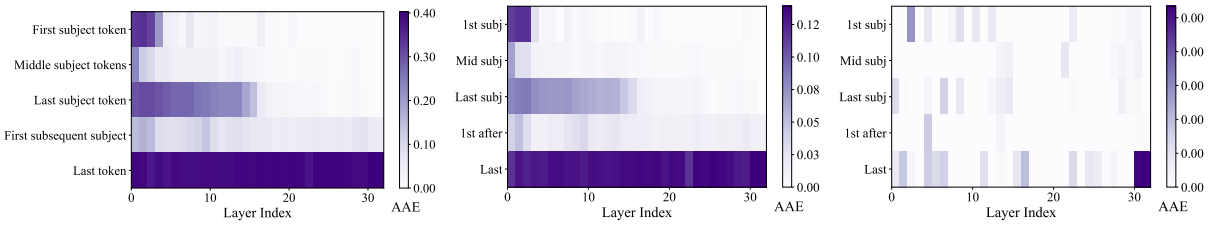
(2) Zeroing Ablation (Necessity). We perform zeroing ablation to test node necessity (Heimersheim and Nanda, 2024). We set activations at specific positions $h^{(l,t)}$ to zero to identify critical nodes. A sharp drop in the probability of the correct answer indicates that the ablated state is necessary for the computation.

Repair Results (Sufficiency). Figure 4 displays the impact of patching clean signals. The 4-bit model (Fig. 4a) shows clear hotspots at the last subject token in early layers. This position is critical for accessing factual knowledge (Meng et al., 2022). Injecting clean signals here significantly restores prediction performance, proving the connection to the final output is intact. In contrast, the 2-bit model (Fig. 4b) is unresponsive to subject patching. This implies that the computational pathway is broken. The layers fail to pass the information forward, even given correct inputs.

Ablation Results (Necessity). Figure 5 identifies the critical causal dependencies. The 4-bit model (Fig. 5b) closely mirrors the FP16 baseline (Fig. 5a), relying on the same last subject token and layers even when the final prediction fails. The reduced intensity suggests that these states are less precise but still functionally necessary. Conversely, the 2-bit model (Fig. 5c) exhibits a diffuse and unstructured pattern. It loses the concentrated critical nodes seen in FP16. This absence of identifiable dependencies indicates a breakdown of the information processing.

Hypothesis Formulation. Combining the macroscopic (Sec. 3.2), layer-wise (Sec. 3.3), and causal (Sec. 3.4) evidence, we formulate the two failure modes hypothesis:

- **Failure Mode I: Signal Degradation.** The model's computational patterns remain largely intact. Quantization error acts as cumulative noise that impairs information precision.
- **Failure Mode II: Computation Collapse.** The quantization error is severe enough to fundamentally damage the functionality of key components. Information cannot be processed correctly and is completely destroyed in the early layers.



(a) FP Model (b) 4-bit Model (c) 2-bit Model

Figure 5: Zeroing ablation analysis on the Failure Subset. The heatmap values represent the Average Ablation Effect (AAE), defined as the decrease in the correct token's prediction probability.

4 Mechanistic Validation and Targeted Intervention

4.1 Analysis of Component-level Impairment

4.1.1 Attention Patterns

A functional attention mechanism should be both focused and accurate; we verify this with normalized attention entropy and focus divergence.

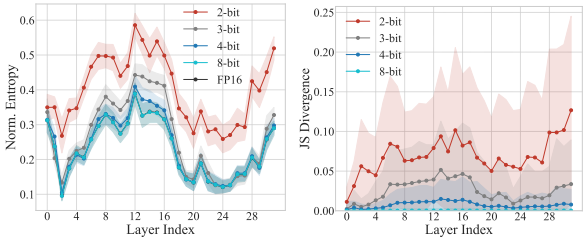
Global Concentration (Entropy). First, we measure the normalized attention entropy to assess if the model can concentrate its attention. For an attention head h at token t, we calculate its Shannon entropy $H(A_{h,t})$ and normalize it by the maximum possible entropy: $E_{norm}(h,t) = H(A_{h,t})/\log_2(t+1)$. We average this across all heads to detect systematic uncertainty.

Focus Divergence (JSD). Entropy alone is insufficient because a model might confidently focus on the wrong token. To measure this deviation, we calculate the Jensen–Shannon divergence (JSD) between the quantized attention distribution (P_Q) and the FP16 baseline (P_{FP}) at the critical last subject token: $JSD(P_{FP}||P_Q) = \frac{1}{2}D_{KL}(P_{FP}||M) + \frac{1}{2}D_{KL}(P_Q||M)$, where M is the average distribution. A high JSD indicates the focus has shifted significantly.

As illustrated in Figure 6, the 4-bit model generally follows the FP16 trend with only slightly increased entropy. In contrast, the 2-bit model exhibits high entropy across all layers, indicating a global failure to concentrate. Meanwhile, its JSD surges significantly, proving that the attention focus deviates fundamentally.

4.1.2 FFN Key-Value Memory

FFN layers function as key-value memories (Geva et al., 2021). For Llama models, the intermediate activation $h_{key} = \text{SiLU}(W_{gate}x) \odot (W_{up}x)$ acts as the "key" to select specific expert neurons. We



(a) Global Entropy (b) Focus Divergence

Figure 6: Analysis of attention mechanisms on the Failure Subset. (a) Normalized Attention Entropy (all tokens). (b) Jensen-Shannon Divergence from the FP16 baseline (last subject token).

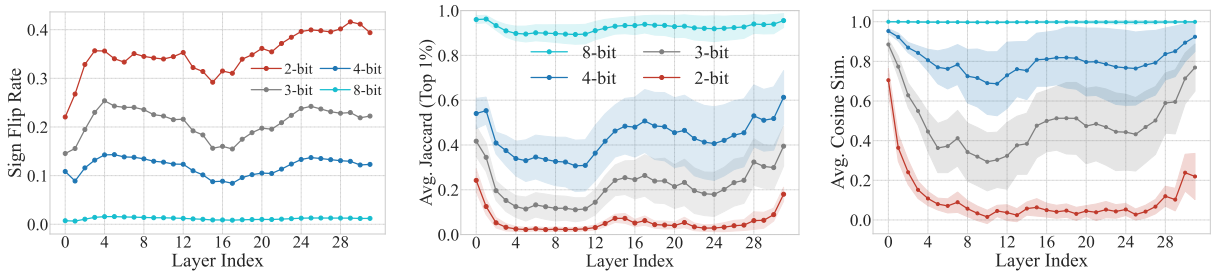
examine the integrity of this key at the last subject token with two metrics.

Gating Consistency (Sign Flip Rate). First, we measure the sign flip rate (SFR) of the gate input ($W_{gate}x$). Since the SwiGLU activation depends on the sign, a noise-caused flip ($\text{sign}(x_Q) \neq \text{sign}(x_{FP})$) can fundamentally reverse the neuron's logical state (active vs. suppressed).

Retrieval Accuracy (Jaccard Index). Second, we use the Jaccard index to check the Top-1% activated neurons in h_{key}. This measures if the model activates the same neurons as the FP16.

As shown in Figure 7a, the 2-bit model exhibits a high sign flip rate ($> 30\%$), indicating quantization noise is large enough to reverse the gate direction. Consequently, the Jaccard Index drops to ≈ 0.1 (Figure 7b), confirming the model activates the wrong neurons. In contrast, the 4-bit model maintains high gating consistency and retrieval overlap.

Analysis of Values (Semantic Direction). Finally, we check the output quality by measuring the cosine similarity between the quantized FFN output ($h_{value} = W_{down}h_{key}$) and the FP16 baseline. This tells us if the retrieved information has the correct semantic direction.



(a) Gate Sign Flip Rate (b) Expert Jaccard Similarity (Top-1%) (c) Value Similarity (Cosine)

Figure 7: Analysis of FFN Key-Value Memory at the last subject token on the Failure Subset. The 2-bit exhibits high gating instability (a) and low expert overlap (b), leading to semantic collapse (c).

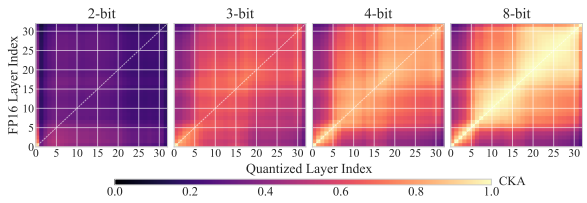


Figure 8: CKA heatmaps of hidden states at the last subject token.

Figure 7c confirms the contrast. The 4-bit model maintains high similarity (≈ 0.8) even when it fails, implying it retrieves the correct concept but with precision errors. In contrast, the 2-bit model drops to near-zero immediately, confirming the retrieved information is completely unrelated to the target. Similar patterns were observed on the Robust Subset, see Appendix B.1.

4.2 Analysis of Representation-level Deviation

Building on the component-level findings, we now examine whether the quantization noise merely blurs the signal or fundamentally destroys the structural integrity of the representation space.

4.2.1 Analysis of Representational Topology

We employ linear centered kernel alignment (CKA) (Kornblith et al., 2019) to analyze the structural correspondence between the activation matrices of quantized and FP16 models.

Figure 8 visualizes results at the last subject token, the critical site for knowledge extraction (Meng et al., 2022) validated by our earlier causal tracing. The diagonal line represents layer-wise correspondence, indicating behavioral similarity to the FP16 model at the same layer. We observe a sharp contrast between the two failure modes. The 4-bit model retains a bright diagonal and block structure similar to 8-bit, only with slightly reduced intensity. This confirms that the global representational structure is preserved. Conversely, the 2-bit model appears almost entirely dark purple. The absence of diagonal structure indicates a "Structural Collapse," where the representational spaces are totally different. Component-wise breakdowns and positional validation are shown in Appendix B.2.

4.2.2 Analysis of Semantic Subspace

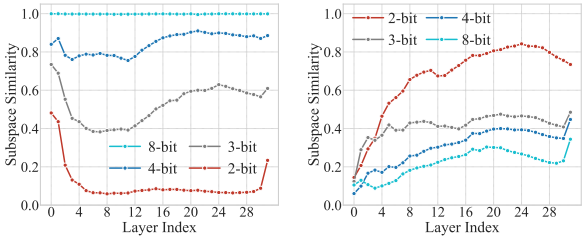
While CKA analyzes the global topology, we use singular value decomposition (SVD) to inspect the internal structure of the activation matrices (A). We conduct two complementary analyses on the Failure Subset.

Activation Subspace Alignment. First, we check if the quantized models utilize the same semantic directions as the FP16 model. We compare the top-k principal directions (columns of V, where $A = USV^T$). We set $k = 50$ (capturing $> 90\%$ of spectral energy) to isolate core semantics from long-tail noise. Let $V_{fp,k}$ and $V_{q,k}$ be the subspaces of the FP16 and quantized models. We calculate their similarity as:

$$Sim(V_{fp}, V_q) = \frac{1}{k} \sum_{i=1}^{k} \sigma_i (V_{fp,k}^T V_{q,k})^2 \quad (1)$$

Figure 9(a) shows that the 4-bit model maintains high similarity (> 0.8) to FP16, confirming its core computational directions remain largely intact even when the model fails. In contrast, the 2-bit model drops to near-zero similarity, indicating a complete loss of the original semantic directions.

Error Subspace Analysis. While activation subspace analysis confirms the deviation of representation directions, it doesn't explain whether the error aligns with the original signal. Consequently, we decompose the error matrix ($E = A_q - A_{fp}$) and measure the alignment between principal error directions (V_{err}) and original signal directions (V_{fp}).



(a) Activation Subspace (b) Error Subspace

Figure 9: Layer-wise SVD analysis (Top-50 dimensions) on the Failure Subset. (a) Similarity of activation subspaces to FP16. (b) Alignment between quantization error and FP16 subspaces.

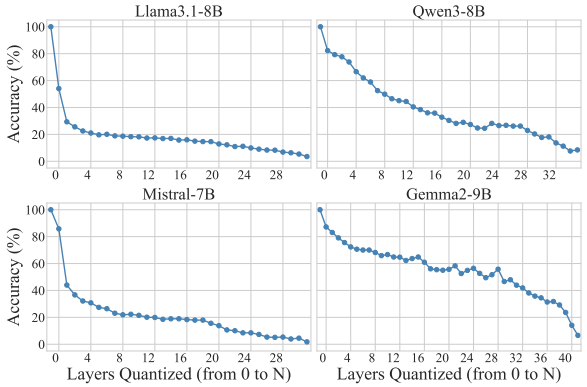


Figure 10: Progressive 4-bit quantization ("domino effect") analysis on the Failure Subset.

Figure 9(b) reveals a critical difference. The 2-bit error is highly aligned with the signal subspace (similarity ≈ 0.8). This means the quantization error is not random noise but directly interferes with the model's primary features. Conversely, the 4-bit error is much less alignment (≈ 0.3), resembling random noise that affects precision without destroying signal structure. Results on the Robust Subset are consistent and shown in Appendix B.2.

Summary of Diagnosis. Combining the component-level and representation-level evidence, we confirm the existence of two distinct failure modes. The 4-bit models exhibit Signal Degradation, where representations are impaired but structurally intact. Conversely, standard 2-bit models exemplify Computation Collapse, where both component functionality and semantic structure are fundamentally destroyed. Crucially, these failures are not strictly tied to specific bit-widths, but reflect the distinct nature of the damage.

4.3 Mechanism-Aware Interventions

Guided by the mechanistic diagnosis, we now demonstrate that the Signal Degradation mode (typical in 4-bit) is localizable and repairable, whereas the Computation Collapse mode (observed in 2-bit) proves systemic and irreversible without retraining.

4.3.1 Signal Degradation: Localization and Repair

The Signal Degradation hypothesis implies that the impairment is not structural but cumulative. We validate this by locating the degradation source and designing a targeted repair.

Localization: The "First Domino" Test. To locate failure origins, we conduct a "domino effect" experiment by progressively quantizing the model from layer 0 to k in 4-bit, keeping subsequent layers in FP16. Figure 10 reveals two distinct, architecture-dependent degradation patterns: **(1) Early Representation Bottleneck** (Llama3.1, Mistral): Accuracy drops sharply when quantizing only the first few layers. **(2) Uniform Degradation** (Qwen3, Gemma2): Performance declines smoothly across all layers. Complementary single-layer quantization and component-level sensitivity analysis are provided in Appendix C.1.

Intervention: A Two-Stage Repair Strategy. Guided by the localization, we design a two-stage intervention to recover the degraded signal.

(1) Source Protection. We first apply targeted protection to mitigate error at its primary sources. For Llama/Mistral, we apply early-layer protection, retaining the first two layers in 8-bit (4.25 avg. bits). For Qwen/Gemma, where sensitivity is distributed, we apply kurtosis-based protection (4.1 avg. bits), preserving high-kurtosis weights that are most vulnerable. This aligns with mixed-precision methods like SPQR (Dettmers et al., 2023), which keep sensitive weights in high precision. It supports our diagnosis that protecting critical components effectively prevents degradation.

Figure 11 (dashed orange) shows this basic protection improves internal signal quality over the baseline (gray). However, final-layer accuracy still lags as cumulative errors weaken the signal until it is surpassed by linguistic noise.

(2) Signal Restoration. To counteract the late-stage competition failure, we introduce peak signal amplification. We identify the layer with the highest confidence (lowest entropy) and amplify its output logits by a factor $\alpha > 1$. As shown in Figure 11 (solid orange), this corrects the late-stage drop and restores the trajectory close to FP16.

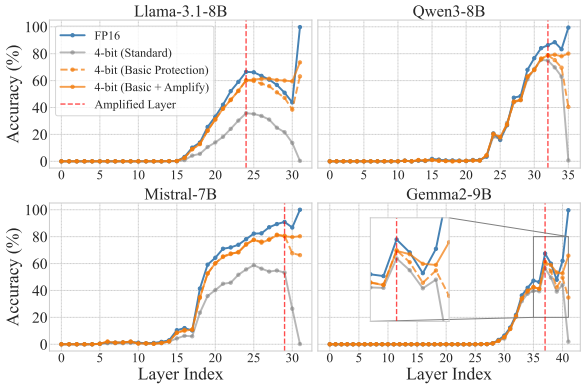


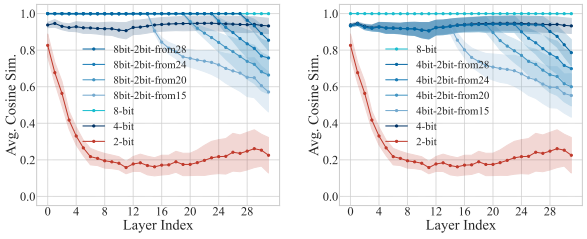
Figure 11: Logit Lens accuracy on the Failure Subset. Our two-stage strategy (orange lines) restores the degraded baseline toward FP16.

Model	Baseline (4-bit)	+ Basic Repair	+ Basic & Amplify (Final)
Llama3.1-8B	0.00%	67.91%	**75.19%** (α=3)
Mistral-7B	0.00%	66.86%	**81.26%** (α=9)
Qwen3-8B	0.00%	40.24%	**79.88%** (α=7)
Gemma2-9B	0.00%	33.85%	**64.08%** (α=2)

Table 1: 4-bit intervention results on the Failure Subset.

Quantized Layers	Accuracy on Subsets (%) Robust	Failure
None (FP16)	100.00	100.00
$k = 0$ (Layer 0)	65.47	15.03
$k = 1$ (Layers 0-1)	41.65	5.29
$k = 2$ (Layers 0-2)	24.66	2.50
$k = 3$ (Layers 0-3)	10.72	1.04
$k = 5$ (Layers 0-5)	2.51	0.38

Table 2: The "domino effect" of 2-bit damage on Llama3.1-8B. Models are quantized from layer 0 to k in 2-bit, with subsequent layers remaining FP16.



(a) 8-bit to 2-bit (b) 4-bit to 2-bit

Figure 12: Layer output cosine similarity under high-precision signal injection on the Robust Subset.

As summarized in Table 1, this combined strategy yields substantial gains across all models, confirming that 4-bit failure is a recoverable impairment of signal intensity.

4.3.2 Computation Collapse: Systemic Irreversibility

In contrast, we posit that Computation Collapse is a systemic processing failure. We validate its irreversibility under training-free interventions through three complementary analyses.

(1) Irreversibility of Damage. We apply the same "domino" test to 2-bit models. Table 2 shows catastrophic results: for Llama3, quantizing just the first two layers ($k = 1$) causes accuracy to plummet from 100% to 41.65%. This proves that 2-bit damage is instantaneous and irreversible, where the signal is destroyed at the source, and even 30 subsequent FP16 layers cannot recover it.

(2) Failure to Process High-Precision Signals. We further test if 2-bit components can function when provided with high-quality signal. We keep the first k layers at high precision (8/4-bit) and quantize subsequent layers to 2-bit. As Figure 12 shows, cosine similarity remains high (> 0.9) initially but collapses immediately upon entering the 2-bit layers. This confirms 2-bit components are computationally non-functional, failing to sustain information even given perfect input. Component-level analysis is shown in Appendix C.

(3) Failure of Mere Compensation. We attempt to recover performance using both our protection strategies (highly effective against Signal Degradation) and EORA (Liu et al., 2024), an advanced low-rank compensation method. However, the 2-bit collapse resists all such interventions. This confirms that the failure stems from a fundamental component malfunction rather than localized precision loss, necessitating structural reconstruction (e.g., fine-tuning) rather than mere compensation.

5 Conclusion

In this work, we bridge the macroscopic performance cliff with microscopic mechanistic failures. We propose and validate the Two Failure Modes Hypothesis, distinguishing between Signal Degradation and Computation Collapse. Our analysis reveals a qualitative shift from Signal Degradation (impaired but functional) to Computation Collapse (fundamental component malfunction). Crucially, the distinct repairability of these modes implies that the collapse necessitates reconstructing computational functionality rather than simple compensation. This work offers a diagnostic foundation for future principled quantization.

Limitations

Our investigation currently focuses on weight-only quantization across representative model families. Consequently, extending these findings to other paradigms, such as activation quantization, remains a direction for future work. Additionally, our evaluation anchors on factual knowledge recall; how the identified failure modes manifest in complex reasoning tasks deserves separate investigation.

Acknowledgments

This work was supported by Beijing Natural Science Foundation (L243006), the National Natural Science Foundation of China (No.62406321), the independent research project of the Key Laboratory of Cognition and Decision Intelligence for Complex Systems and CIPS-SMP-Zhipu Large Model Fund.

References

Ting-Yun Chang, Muru Zhang, Jesse Thomason, and Robin Jia. 2025. Why Do Some Inputs Break Low-Bit LLM Quantization? arXiv.

Vishnu Kabir Chhabra and Mohammad Mahdi Khalili. 2025. Towards Understanding and Improving Refusal in Compressed Models via Mechanistic Interpretability. *arXiv preprint*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint*. ArXiv:2110.14168 [cs].

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. 2023. SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression. arXiv.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2020. HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 18518–18529. Curran Associates, Inc.

Razvan-Gabriel Dumitru, Vikas Yadav, Rishabh Maheshwary, Paul Ioan Clotan, Sathwik Tejaswi Madhusudhan, and Mihai Surdeanu. 2025. Variable Layerwise Quantization: A Simple and Effective Approach to Quantize LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 534–550, Vienna, Austria. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. In *The International Conference on Learning Representations*.

Yao Fu, Xianxuan Long, Runchao Li, Haotian Yu, Mu Sheng, Xiaotian Han, Yu Yin, and Pan Li. 2025. Quantized but Deceptive? A Multi-Dimensional Truthfulness Evaluation of Quantized LLMs. arXiv. ArXiv:2508.19432 [cs].

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. arXiv.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 181 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ADS Bibcode: 2024arXiv240721783G.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*. ArXiv:2310.06825 [cs].

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. arXiv.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR. ISSN: 2640-3498.

Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Evaluating Quantized Large Language Models. arXiv.

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.

Ruikang Liu, Yuxuan Sun, Manyi Zhang, Haoli Bai, Xianzhi Yu, Tiezheng Yu, Chun Yuan, and Lu Hou. 2025a. Quantization Hurts Reasoning? An Empirical Study on Quantized Reasoning Models. arXiv.

Shih-Yang Liu, Maksim Khadkevich, Nai Chit Fung, Charbel Sakr, Chao-Han Huck Yang, Chien-Yi Wang, Saurav Muralidharan, Hongxu Yin, Kwang-Ting Cheng, Jan Kautz, Yu-Chiang Frank Wang, Pavlo Molchanov, and Min-Hung Chen. 2024. EoRA: Fine-tuning-free Compensation for Compressed LLM with Eigenspace Low-Rank Approximation. *arXiv preprint*.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2025b. SpinQuant: LLM quantization with learned rotations. arXiv. ArXiv:2405.16406 [cs].

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

Satya Sai Srinath Namburi, Makesh Sreedhar, Srinath Srinivasan, and Frederic Sala. 2023. Investigating the Impact of Compression on Parametric Knowledge in Language Models.

nostalgebraist. 2020. interpreting GPT: the logit lens.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint*. ArXiv:2408.00118 [cs].

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. 2024. QuIP#: Even Better LLM Quantization with Hadamard Incoherence and Lattice Codebooks. *arXiv preprint*. ArXiv:2402.04396 [cs].

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 *Proceedings of the 40th International Conference on Machine Learning*, pages 38087–38099. PMLR.

He Xiao, Qingyao Yang, Dirui Xie, Wendong Xu, Wenyong Zhou, Haobo Liu, Zhengwu Liu, and Ngai Wong. 2025. Exploring Layer-wise Information Effectiveness for Post-Training Quantization in Small Language Models. *arXiv preprint*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Feng Zhang, Yanbin Liu, Weihua Li, Jie Lv, Xiaodan Wang, and Quan Bai. 2025a. Towards Superior Quantization Accuracy: A Layer-sensitive Approach. *arXiv preprint*.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. 2025b. Catastrophic Failure of LLM Unlearning via Quantization. arXiv.

A Experimental Details

A.1 Quantization Configuration

We use GPTQModel for post-training quantization with group size 128. Calibration is performed on 128 randomly sampled C4 sequences of length 2048 (Raffel et al., 2020). All subsequent evaluations use greedy decoding (temperature = 0) to ensure deterministic inference.

A.2 Prompt Templates

Primary Templates (Mechanistic Analysis). For the primary mechanistic analysis, we select one specific template per relation that naturally ends with the object to facilitate next-token probing. Table 3 provides examples of the templates used for different relation types.

Relation ID	Template
P19 (Place of Birth)	[X] was born in
P27 (Country of Citizenship)	[X] is a citizen of
P36 (Capital)	The capital of [X] is
P106 (Profession)	The profession of [X] is

Table 3: Examples of standardized templates used for mechanistic analysis.

Robustness Templates (Phenomenological Check). For the robustness evaluation in Figure 1, we utilize the full set of Pararel paraphrases. To handle varying target positions [Y] across patterns (e.g., "[X]'s capital is [Y]", "[Y] is the capital of [X]"), we standardize the input by wrapping statements into an instruction: *Based on your knowledge, complete the following sentence by filling in the blank: '{cloze_statement}' The missing word is:* This ensures the model generates the target entity as the immediate completion, regardless of the original sentence structure.

A.3 Dataset Partition Statistics

Table 4 details the sample counts for the Robust Subset (`fp_and_4bit_correct`) and the Failure Subset (`fp_correct_4bit_wrong`) across all evaluated models.

Model	Total	Robust	Failure
Llama-3.1-8B	7,777	5,661	2,116
Qwen3-8B	4,055	3,558	497
Mistral-7B	4,294	3,787	507
Gemma-2-9B	6,375	5,601	774

Table 4: Sample counts for analysis subsets.

B Supplementary Mechanistic Validation

B.1 Component-level Impairment

Attention Pattern (Entropy & JSD). Figures 13 and 14 confirm that the high uncertainty and attention divergence in 2-bit models are universal across datasets and token positions. Notably, while 4-bit models show tighter alignment on the Robust Subset (Fig. 14a), 2-bit models consistently exhibit significant divergence.

FFN Key-Value Memory. Figures 15 and 16 confirm that the collapse in 2-bit is universal, regardless of task difficulty or token position. Specifically, 2-bit models consistently exhibit extreme sign flip rates and near-zero Jaccard scores (Panels a & b), indicating a complete breakdown in expert selection. This leads to a semantic collapse in the Value outputs (Panel c), where similarity drops to near-zero. In contrast, 4-bit models maintain strong alignment, exhibiting higher mean similarity and lower variance compared to the difficult subset.

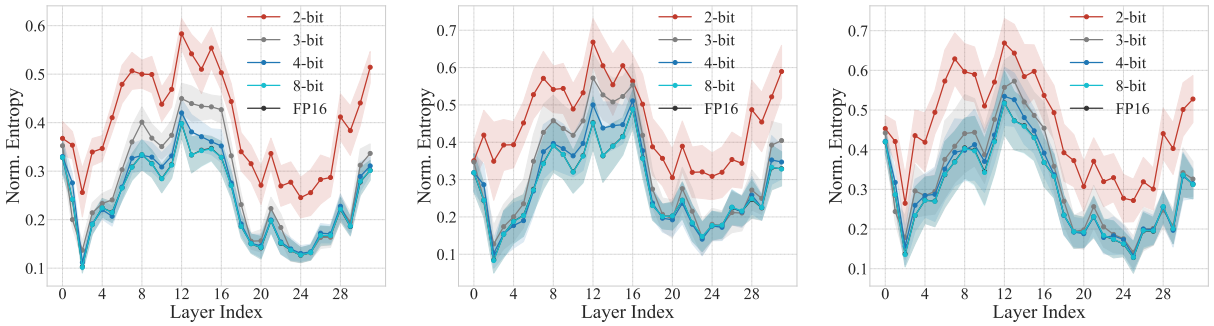
B.2 Representational Topology

CKA (Components & Position). We expand the CKA analysis to specific components and different token positions. Figure 17 analyzes the internal components at the last subject token. It shows that while the layer output retains some structure due to residual connections, the internal components of the 2-bit model are completely collapsed (pitch black). Figure 18 repeats the analysis at the last token. The trend remains identical: 4-bit models preserve the high-correlation block structure, while 2-bit models lose structural coherence.

Semantic Direction (Cosine Similarity). While the main text analyzes the internal structure at the subject token, here we utilize cosine similarity at the last token to verify the ultimate output of the representation.

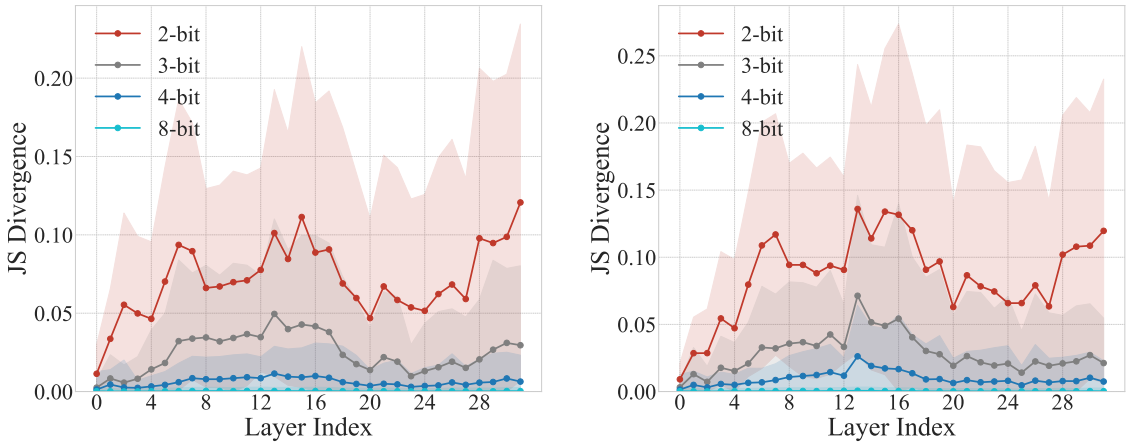
Figure 19 compares the layer output similarity. The 2-bit model suffers a complete collapse, with similarity dropping to near-zero. The 4-bit model maintains high alignment. However, on the failure subset (Fig. 19b), it shows larger variance compared to the success subset (Fig. 19a). This suggests 4-bit failures come from noise instability rather than directional error.

SVD Analysis. Figure 21 presents the comparative SVD analysis on the Robust Subset to verify the consistency of our findings on easier samples.



(a) All tokens on the Robust Subset (b) Last subject token on the Failure (c) Last token on the Failure Subset

Figure 13: Supplementary results for Normalized Attention Entropy.



(a) Last subject token on the Robust Subset (b) Last token on the Failure Subset

Figure 14: Supplementary JSD Analysis. (a) 4-bit models maintain high alignment on the Robust Subset, while 2-bit models show instability. (b) Divergence persists at the last token.

C Intervention and Sensitivity Analysis

C.1 Localized Sensitivity in 4-bit Models

Layer-wise Sensitivity. Figure 22 complements the main text's "domino" analysis. We quantize only a single layer to 4-bit while keeping others in FP16. The results confirm the architecture-dependent sensitivity: Llama/Mistral show extreme sensitivity in early layers, while Qwen/Gemma show uniform sensitivity.

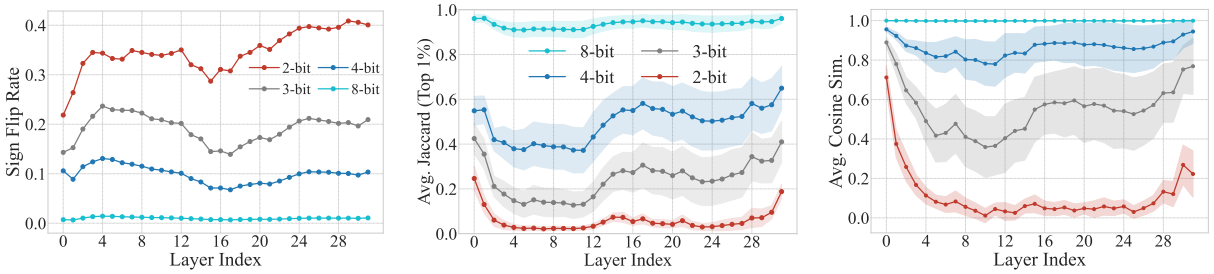
Component-wise Sensitivity. We analyze the sensitivity of individual components by quantizing them separately to 4-bit (Table 5).

- **Localized Vulnerability (Llama/Mistral)**: MLP modules are significantly more fragile than Attention modules. Specifically, the "content generation" weights (`down_proj`, `v_proj`) are far more critical than the "routing" weights.
- **Balanced Sensitivity (Qwen/Gemma)**: Degradation is uniform across MLP and Attention modules, with no single component acting as a distinct failure point.

Unlike 4-bit, where some modules remain functional, 2-bit quantization causes a universal failure. No module remains functionally robust, confirming that the failure is driven by a systemic breakdown of representational capacity rather than specific component weak points.

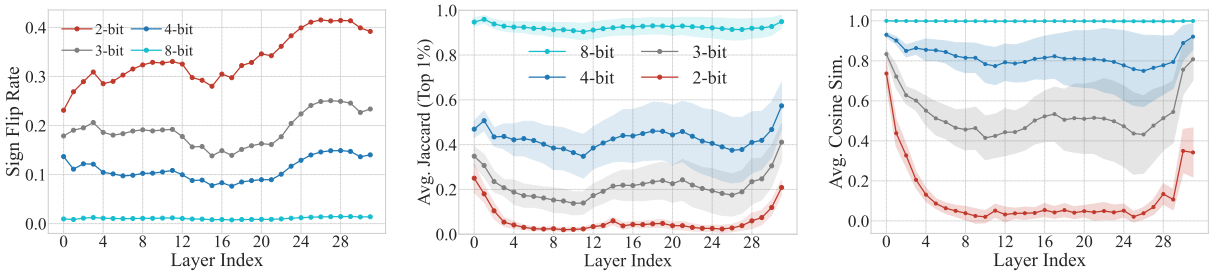
C.2 Systemic Collapse in 2-bit Models

Figure 23 shows single-layer 2-bit quantization results. Unlike 4-bit, quantizing even a single early layer (especially in Llama/Mistral) leads to catastrophic drops. Figure 24 decomposes the signal injection analysis. It confirms that the collapse observed in the main text (Figure 12) occurs simultaneously in both Attention and MLP outputs, proving the failure is systemic.



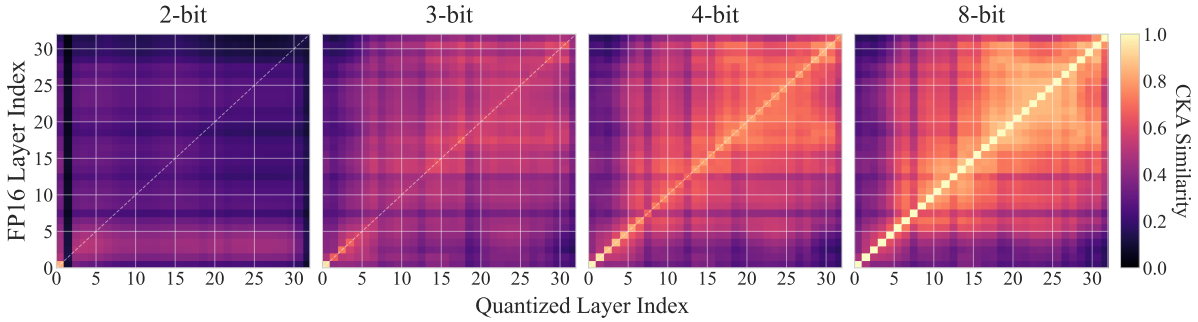
(a) Gate Sign Flip Rate (b) Expert Jaccard Similarity (c) Value Similarity (Cosine)

Figure 15: Supplementary FFN Analysis on the Robust Subset at the last subject token. Even on easier samples, 2-bit models show internal instability (a, b) and output degradation (c), while 4-bit models remain healthy.

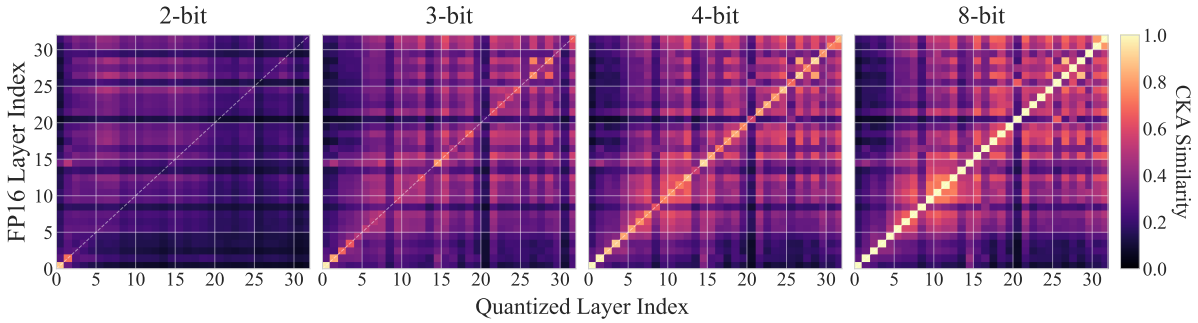


(a) Gate Sign Flip Rate (b) Expert Jaccard Similarity (c) Value Similarity (Cosine)

Figure 16: Supplementary FFN Analysis at the last token on the Failure Subset. The failure mode is consistent across positions: 2-bit causes gating collapse (a) and retrieval failure (b), destroying the final representation (c).

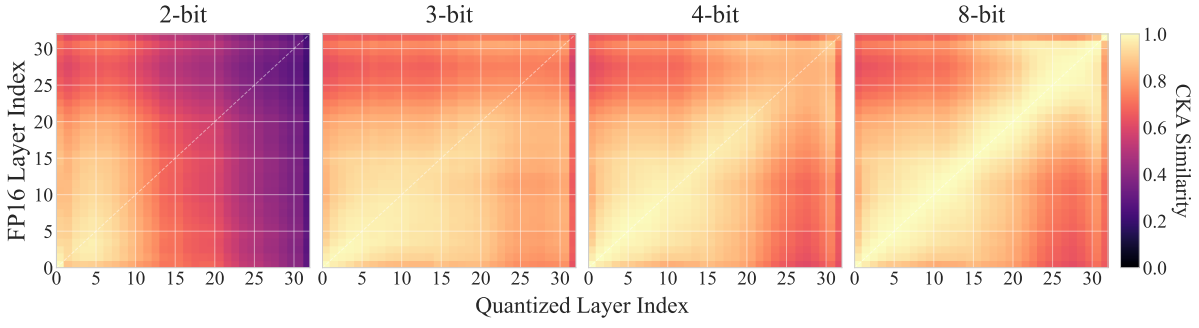


(a) FFN Output (Last Subject Token)

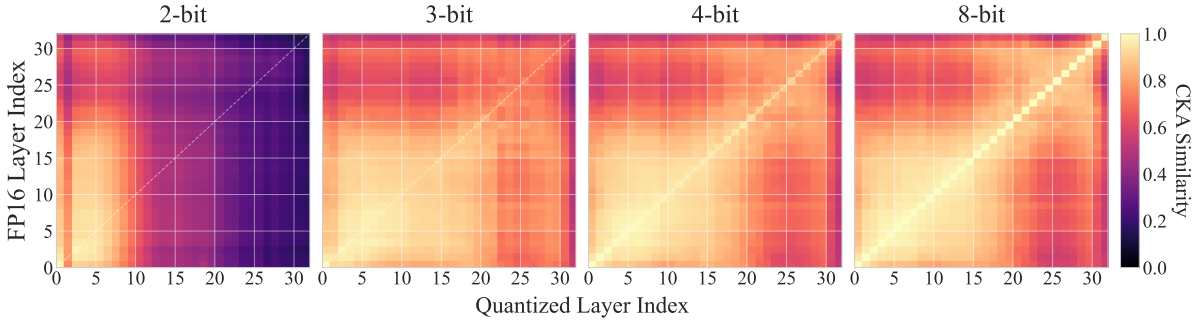


(b) Attention Output (Last Subject Token)

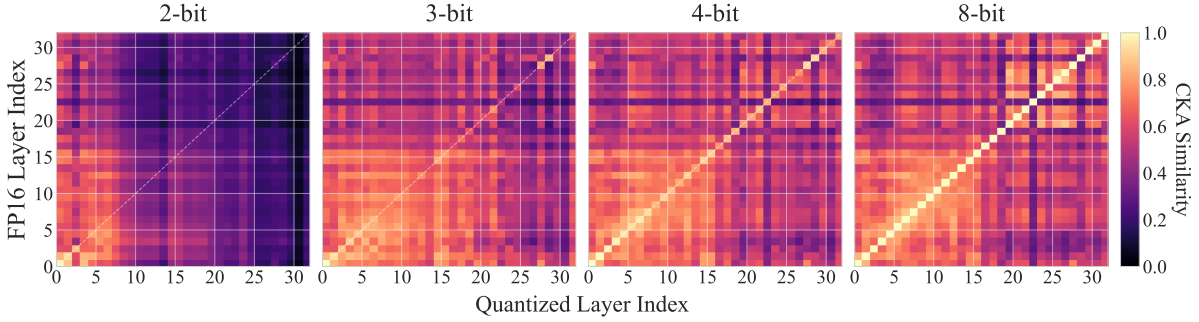
Figure 17: Component-wise CKA Analysis at the last subject token. The figures are stacked vertically to show the detail of FFN and Attention collapse in 2-bit models.



(a) Layer Output (Last Token)

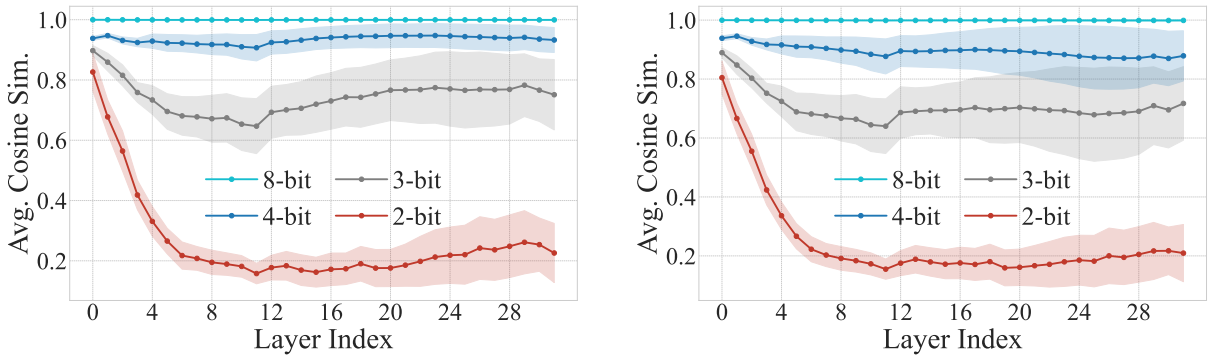


(b) FFN Output (Last Token)



(c) Attention Output (Last Token)

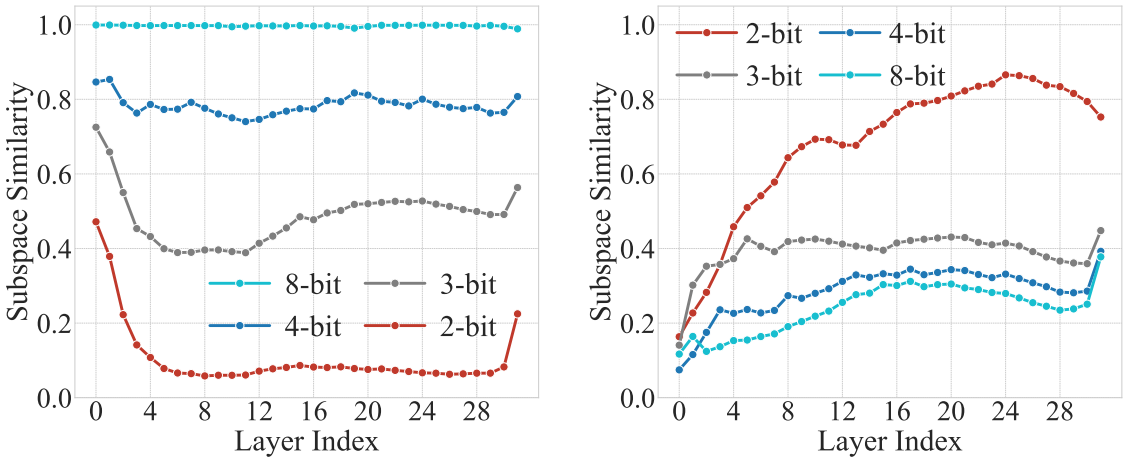
Figure 18: CKA Analysis at the last token. The topological collapse is consistent across all components.



(a) Layer output on the Robust Subset

(b) Layer output on the Failure Subset

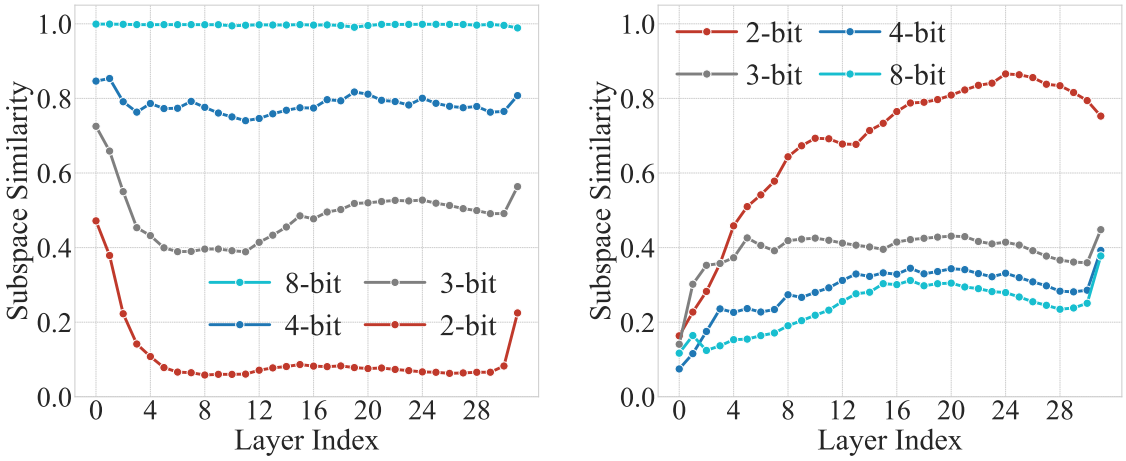
Figure 19: Supplementary Cosine Similarity Analysis at the last token. Comparisons show that while 2-bit models collapse universally, 4-bit models only suffer from instability on difficult samples.



(a) Activation Subspace Alignment

(b) Error-Signal Alignment

Figure 20: Supplementary SVD analysis on the Robust Subset. (a) 4-bit models match FP16. (b) 2-bit error remains destructive (high overlap with signal) even on easier samples.



(a) Activation Subspace Alignment

(b) Error-Signal Subspace Alignment

Figure 21: Supplementary SVD analysis on the Robust Subset. (a) Activation subspace alignment remains high for 4-bit, similar to FP16. (b) Error-signal alignment for 4-bit remains low, while 2-bit error remains highly aligned (destructive).

Model Family	**All**	**MLP**	**Attn**	**MLP Gate/Up**	**MLP Down**	**QK Proj**	**V Proj**	**O Proj**
Llama-3.1-8B-2bit	0.00	0.00	9.40	4.82	4.16	48.91	21.12	41.45
Llama-3.1-8B-4bit	0.00	38.04	53.88	70.65	49.57	89.70	54.44	88.61
Mistral-7B-4bit	0.00	43.79	61.93	86.59	45.76	89.94	65.68	89.74
Qwen3-8B-4bit	0.00	49.30	46.88	63.58	64.19	71.03	69.01	78.87
Gemma-2-9B-4bit	0.00	54.13	52.33	50.65	70.03	76.10	61.11	81.78

Table 5: Component-level sensitivity analysis on the Failure Subset. Values denote accuracy (%) when only the specific component is quantized, highlighting the contrast between localized fragility (Llama/Mistral) and balanced sensitivity (Qwen/Gemma).

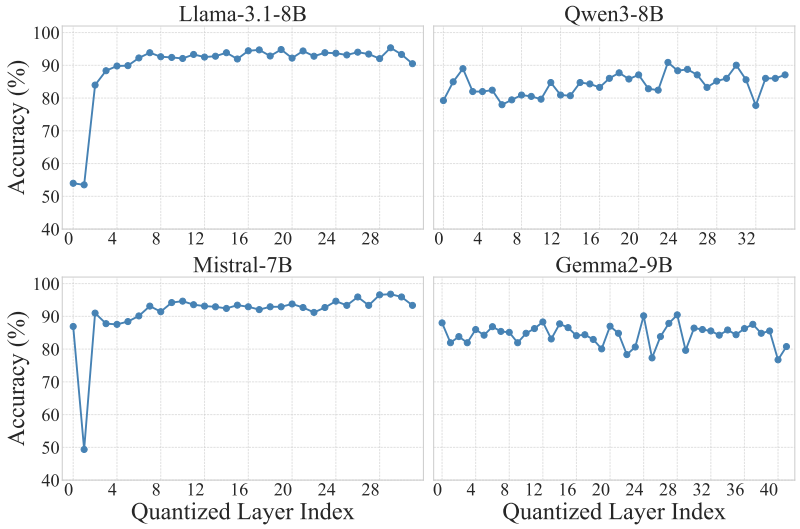


Figure 22: Single-layer 4-bit quantization sensitivity on the Failure Subset. Llama/Mistral show localized fragility, while Qwen/Gemma are balanced.

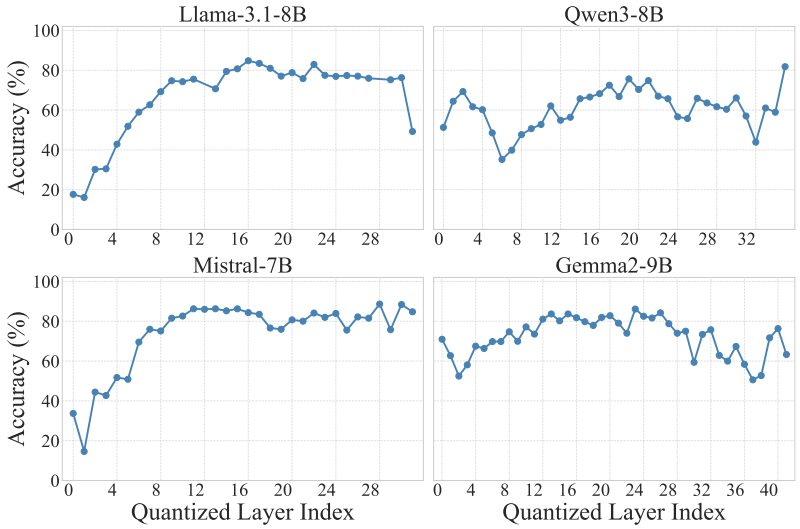
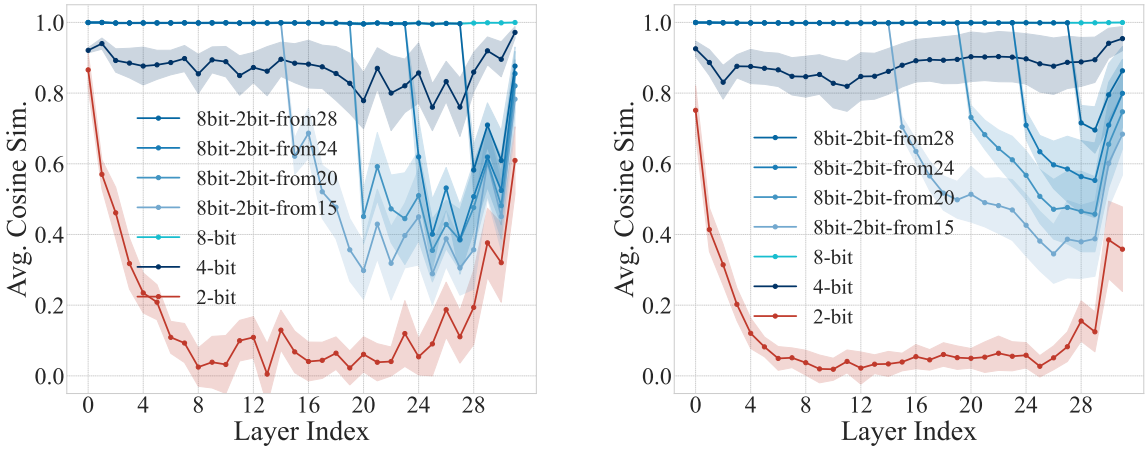


Figure 23: Single-layer 2-bit quantization sensitivity on the Failure Subset. Catastrophic drops from early layers (Llama/Mistral) are evident.



(a) Attention Output

(b) MLP Output

Figure 24: Component decomposition for high-precision signal injection on the Robust Subset. Both Attn and MLP outputs collapse upon entering 2-bit layers.

D Generalizability to AWQ Algorithm

To verify whether our discovered failure modes generalize across quantization algorithms, we replicate the mechanistic analysis using AWQ (Lin et al., 2024) on Llama-3.1-8B. We evaluate the models on the same Failure Subset. The macro-level accuracy strictly mirrors our GPTQ findings: AWQ 4-bit (28.17%) $\rightarrow$ 3-bit (18.01%) $\rightarrow$ 2-bit (0.00%).

D.1 Layer-wise Knowledge Probing

Figure 25 traces the layer-wise knowledge signals. Consistent with GPTQ, the 4-bit and 3-bit AWQ models exhibit Signal Degradation. Their target probabilities build up in deeper layers but remain lower than the FP16 baseline, accompanied by a moderate drop in target ranks. In contrast, the 2-bit model fails to recover any meaningful probability distribution, remaining completely flat at zero across all layers.

D.2 Component-level Impairment

Attention Mechanism. Figure 26 illustrates the attention patterns. While 4-bit and 3-bit models show tight alignment with the baseline, 2-bit quantization triggers a severe concentration collapse. Its normalized attention entropy exceeds 0.80 in middle-to-late layers, and its focus divergence sharply increases, indicating that the attention mechanism loses its routing capability.

FFN Key-Value Memory. Figure 27 presents the FFN functionality metrics. The 4-bit and 3-bit models maintain relatively stable gate flip rates and retrieve semantic values with high cosine similarity. However, the 2-bit model induces massive gate flipping, reaching nearly 80% in middle layers. This severe disruption causes a rapid drop in expert Jaccard similarity and drives the semantic alignment of output values to near-zero.

Collectively, these results confirm that the transition from Signal Degradation to Computation Collapse is a fundamental pattern of quantization damage, rather than a GPTQ-specific artifact.

E Generalizability to Broader Language Tasks

To demonstrate that the discovered failure modes generalize beyond factual recall, we extend our mechanistic metrics to MMLU (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021).

E.1 Experimental Setup

All evaluations are conducted in a 5-shot setup, with full-dataset accuracy reported in Table 6. For the mechanistic analysis, we use the complete GSM8K dataset alongside a representative MMLU subset of 1,066 samples across four diverse domains (macroeconomics, philosophy, clinical knowledge, and computer science).

Dataset	FP16	4-bit	3-bit	2-bit
MMLU	63.42%	61.44%	49.11%	24.18%
GSM8K	56.79%	51.33%	10.16%	0.99%

Table 6: Accuracy of Llama-3.1-8B on broader tasks across bit-widths.

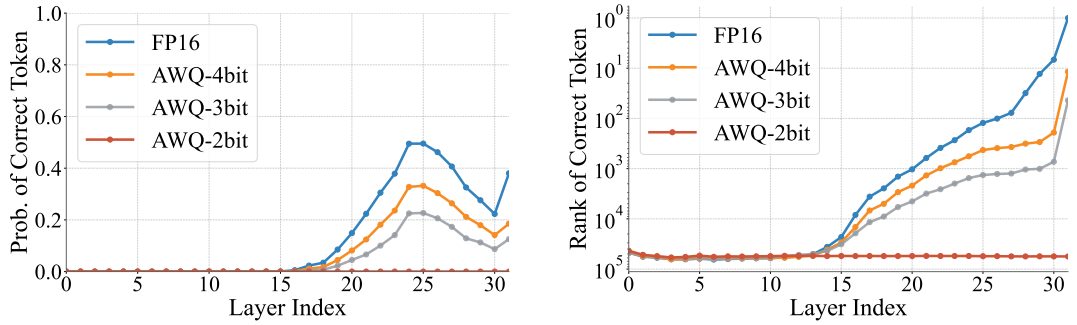
E.2 Semantic Subspace Integrity

We analyze the semantic subspace alignment using a single forward pass for both tasks. Figure 28 illustrates the layer-wise activation subspace similarity. Consistent with our findings on factual recall, the 2-bit trajectory plummets and remains near zero across all layers. In contrast, the 4-bit and 3-bit models initially experience a drop in similarity but subsequently recover and stabilize in deeper layers, confirming that their primary semantic directions are partially preserved despite precision loss.

E.3 Attention and Generation Dynamics

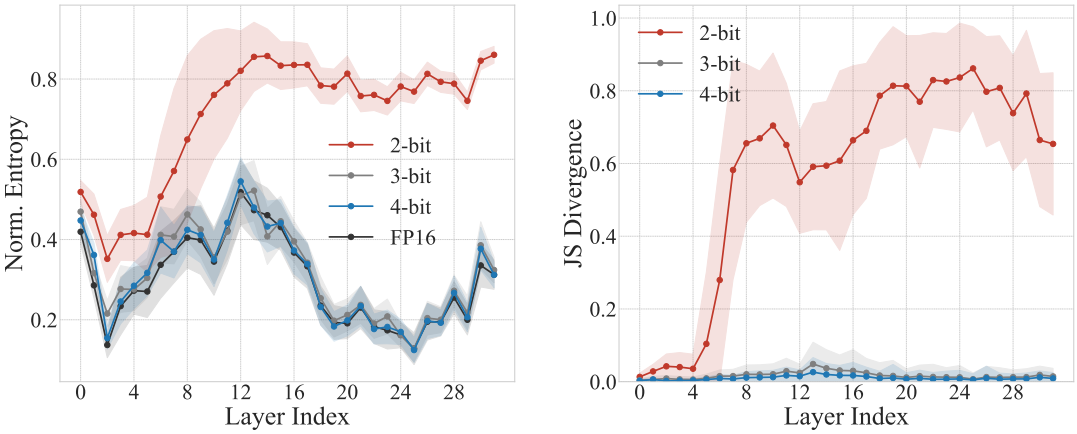
Given the crucial role of the attention mechanism in processing context during multi-step reasoning, attention entropy serves as an effective indicator for observing quantization-induced behavioral shifts. For MMLU (Fig. 29a), the entropy is calculated from a single forward pass. For GSM8K (Fig. 29b), we track the layer-averaged entropy of the last token at each generation step.

As shown in Figure 29, the 2-bit model exhibits a severe deterioration of attention focus. On GSM8K, its attention entropy starts abnormally high at the beginning and persists throughout all steps, whereas the 4-bit entropy closely tracks the FP16 baseline. Because of this persistent high-entropy state, the 2-bit model fails to execute fine-grained reasoning. Qualitative inspection of failure cases reveals it generates chaotic content (e.g., meaningless numbers and repetitive loops), typically failing to halt until hitting the maximum generation length (with median generated tokens doubling from 77 in FP16 to 151 in 2-bit).



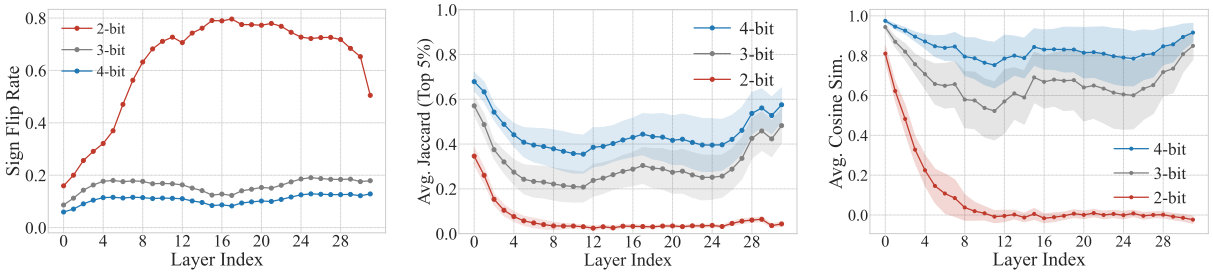
(a) Probability Evolution (b) Rank Evolution

Figure 25: Layer-wise evolution of probability and rank for AWQ on the Failure Subset.



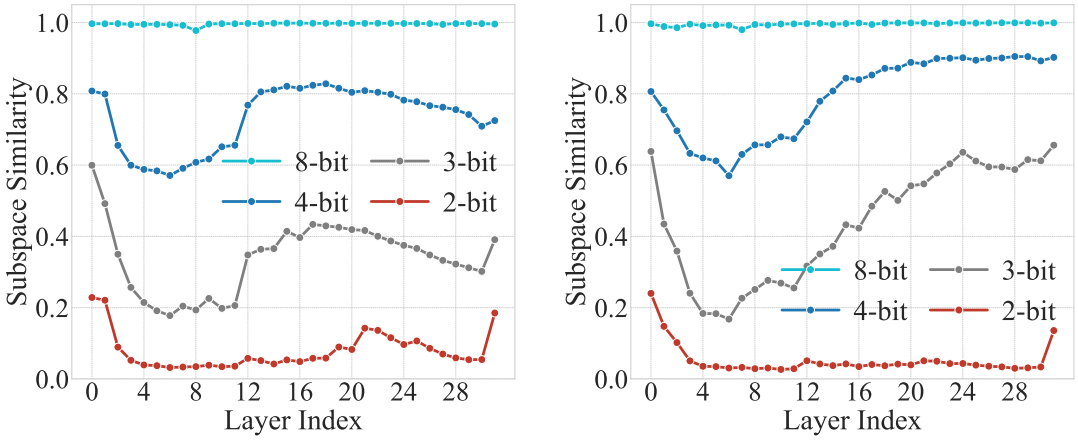
(a) Normalized Attention Entropy (b) Focus Divergence (JSD)

Figure 26: Attention mechanism analysis at the last token for AWQ on the Failure Subset.



(a) Gate Sign Flip Rate (b) Expert Jaccard Similarity (c) Value Similarity (Cosine)

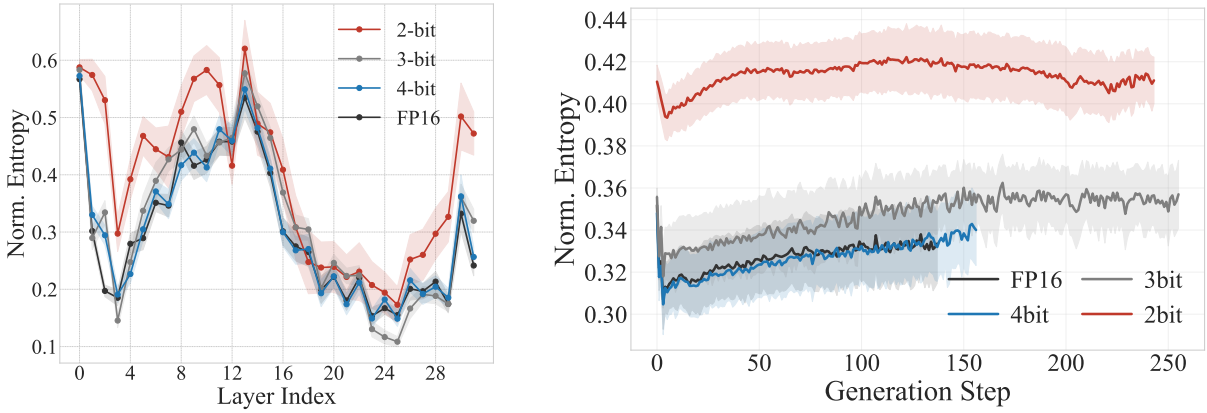
Figure 27: Parallel indicators of FFN functionality at the last subject token for AWQ on the Failure Subset.



(a) Subspace Similarity on MMLU

(b) Subspace Similarity on GSM8K

Figure 28: Layer-wise SVD analysis (Top-50 dimensions) on broader tasks, calculated from a single forward pass.



(a) Layer-wise Entropy on MMLU

(b) Temporal Entropy Dynamics on GSM8K

Figure 29: Attention entropy analysis. MMLU results are calculated from a single forward pass, while the GSM8K curve traces the layer-averaged entropy at each generation step (truncated when <10% of samples remain active).