Residual Geometry as a Diagnostic for Hidden Failure Modes in Fast Language Model Compression

Jun-Sik Yoo
Department of Physics
Korea University
Seoul, Korea, 02846
junsik@korea.ac.kr

Abstract

Low-bit quantization is usually judged by whether the compressed model still works: perplexity, output KL, downstream accuracy, or layer reconstruction error. These metrics are useful, but they can hide which part of the internal computation was preserved. We study 4-bit language-model compression by decomposing each transformer layer update into a teacher-fixed predictable drift and a drift-subtracted residual. On Llama-3.1-8B with NF4 quantization, moving the quantized residual toward the corresponding full-precision teacher residual reduces teacher-to-model output KL, while removing, shuffling, or randomizing the residual sharply worsens it. The same intervention improves teacher-faithful multiple-choice distributions on PIQA and ARC-Easy. Finally, low-frequency modes of a teacher-residual graph recover part of the oracle correction gain, suggesting that fast compression should preserve behaviorally relevant residual-update structure.

1 Introduction

Low-bit quantization is central to fast language-model deployment. It reduces memory footprint and can improve inference throughput when supported by efficient kernels. Scientific workflows increasingly use language models and transformers for code generation, data analysis, simulation control, literature triage, and interactive reasoning. In these settings, compression is attractive, but coarse fidelity metrics may not reveal which internal computations are damaged.

Most post-training quantization methods are evaluated by perplexity, downstream accuracy, output KL, or layer reconstruction error. These metrics answer whether a compressed model remains close to a full-precision model, but they do not identify which component of a layer update was preserved. We ask a more diagnostic question: *what part of the layer-update geometry is damaged by low-bit quantization?*

For hidden state h_l at layer l, define the layer update

$$\Delta h_l = h_{l+1} - h_l. \tag{1}$$

We fit a teacher-fixed map T_l that predicts the dominant update from the current hidden state, and define a drift-subtracted residual

$$\Delta h_l = T_l(h_l) + r_l. \tag{2}$$

The map T_l is a centered ridge map fit only on full-precision teacher activations. We then apply the same fixed map to the quantized model, placing teacher and quantized updates in the same diagnostic coordinate system.

Preprint.

Our main result is that the residual component is behaviorally relevant. At Llama-3.1-8B layer 29 under 4-bit NF4 quantization, replacing the quantized residual by a structured interpolation toward the teacher residual reduces output KL, while removal, token shuffling, and random same-norm controls worsen it. This suggests that the residual is not merely a small leftover hidden by the bulk update: its token-aligned structure matters for behavior.

This work is diagnostic. We do not propose a deployable quantizer. Instead, we identify a failure mode that may guide future mixed-precision, calibration, or correction schemes: limited compression budget should be spent on behaviorally relevant residual-update structure, not only on raw weight or activation error.

Related work. Post-training quantization methods such as LLM.int8, GPTQ, SmoothQuant, NF4/QLoRA, and AWQ show that low-precision inference can preserve high-level model quality when quantization is aware of activation structure, outliers, or approximate reconstruction objectives [Dettmers et al., 2022, Frantar et al., 2023, Xiao et al., 2023, Dettmers et al., 2023, Lin et al., 2024]. Recent work also shows that low-bit quantization errors can be input-dependent and related to late residual stream behavior [Chang et al., 2025]. Our contribution is complementary: we analyze the geometry of the layer update itself and separate a dominant teacher-fixed drift from a residual component that is hidden by aggregate update metrics.

2 Residual-Geometry Diagnostic

For each layer l, we fit a centered ridge map

$$T_l(h) = (h - \mu_h)B_l + \mu_{\Delta h}, \tag{3}$$

where B_l is fit on full-precision teacher activations and layer updates. The teacher residual and quantized residual are

$$r_l^{\mathrm{fp}} = \Delta h_l^{\mathrm{fp}} - T_l(h_l^{\mathrm{fp}}), \qquad r_l^{\mathrm{q}} = \Delta h_l^{\mathrm{q}} - T_l(h_l^{\mathrm{q}}). \tag{4}$$

The residual is an analysis object, not a learned module in the transformer. It measures the component of the layer update not explained by the teacher-fixed predictable drift.

To test whether this residual matters, we intervene during a quantized forward pass. At layer l, we replace the quantized residual by

$$r_l^{\mathrm{patch}}(\alpha) = (1-\alpha)r_l^{\mathrm{q}} + \alpha r_l^{\mathrm{fp}}, \tag{5}$$

and continue the forward pass from

$$h_{l+1}^{\mathrm{patch}} = h_l^{\mathrm{q}} + T_l(h_l^{\mathrm{q}}) + r_l^{\mathrm{patch}}(\alpha). \tag{6}$$

We compare this structured correction with three controls: removing the residual, shuffling teacher residuals across token positions, and replacing residuals by random same-norm vectors. We evaluate teacher-to-model output KL,

$$D_{\mathrm{KL}}(p_{\mathrm{fp}}(\cdot \mid x) \,\|\, p_{\mathrm{patch}}(\cdot \mid x)), \tag{7}$$

averaged over non-padding next-token positions.

3 Experiments and Results

Setup. The main experiment uses Llama-3.1-8B as the full-precision teacher and the same checkpoint loaded with BNB NF4 4-bit quantization as the compressed model. We focus on layer 29 out of 32 blocks: this layer is late enough to affect output distributions, but remains compatible with residual transplantation; see Section 4. The teacher-fixed ridge map is fit on approximately 3.3×10^4 WikiText-2 training token positions. On held-out tokens, the map has validation $R^2 = 0.922$ and update cosine 0.958, capturing the bulk update while leaving a nonzero residual.

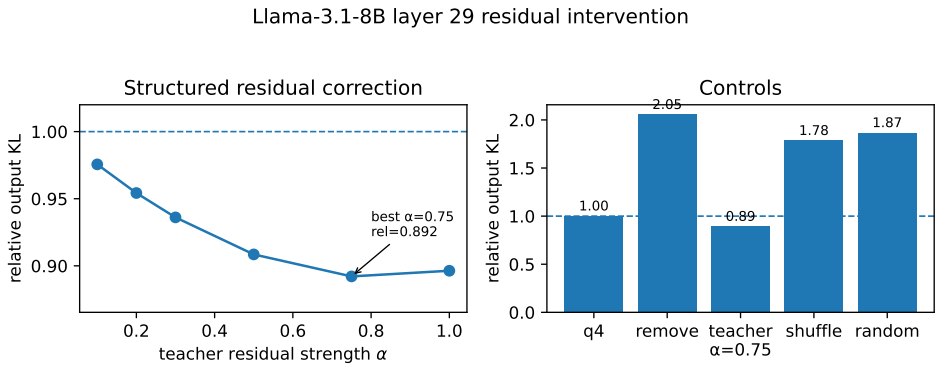

Figure 1: Residual interventions at Llama-3.1-8B layer 29. Structured teacher-residual correction reduces output KL, while residual removal, token shuffling, and random same-norm controls worsen it.

Table 1: Output intervention at Llama-3.1-8B layer 29 under NF4 q4 quantization.

Condition	Mean token KL	Relative to q4
q4 baseline	0.07794	1.000
Remove residual	0.16014	2.055
Teacher residual, $\alpha = 0.30$	0.07295	0.936
Teacher residual, $\alpha = 0.50$	0.07081	0.909
Teacher residual, $\alpha = 0.75$	**0.06953**	**0.892**
Shuffle residual, $\alpha = 0.75$	0.13909	1.785
Random same-norm residual, $\alpha = 0.75$	0.14543	1.866

Residual interventions reduce output KL. Figure 1 and Table 1 show the main intervention result on 128 WikiText-2 validation texts, corresponding to 9,889 evaluated token positions. The q4 baseline has mean token KL 0.07794 to the full-precision teacher. Removing the residual increases KL by a factor of 2.06. In contrast, moving the q4 residual toward the teacher residual reduces KL, with the best fixed intervention at $\alpha = 0.75$ giving KL 0.06953, a relative KL of 0.892. Token-shuffled and random same-norm residual controls at the same strength increase KL to 1.785 and 1.866 times the q4 baseline. Thus the benefit is not explained by residual magnitude alone; it depends on token-aligned residual structure.

Downstream choice distributions also move toward the teacher. We next evaluate multiple-choice probes. For prompt c and answer choices $\{a_j\}$, we compute each choice score as conditional log-likelihood $s_j = \log p(a_j \mid c)$ and compare the resulting softmax choice distribution to the teacher choice distribution using choice KL. This is a teacher-faithfulness probe rather than a leaderboard evaluation.

Table 2 summarizes PIQA and ARC-Easy. On teacher-correct examples, structured residual correction modestly but consistently reduces choice KL, while residual removal strongly worsens it. Boundary subsets, where the full-precision teacher predicts the gold answer but q4 does not, are small but sensitive: on ARC-Easy boundary examples, residual correction reduces choice KL to 0.761 of baseline and recovers some lost decisions. Accuracy changes are modest on the full sets because q4 already agrees with the teacher on many examples.

Recoverable error has collective residual structure. The full teacher-residual intervention is an oracle probe. We therefore ask whether part of the recoverable residual quantization error is organized in a lower-dimensional collective basis. Let

$$\Delta R = R^{\mathrm{fp}} - R^{\mathrm{q}} \tag{8}$$

be the residual quantization error over token positions. We build a k-nearest-neighbor graph on teacher residual directions, form the normalized graph Laplacian, and project ΔR onto the lowest

Table 2: Downstream choice-distribution interventions. We report teacher-to-model choice KL relative to the q4 baseline and accuracy before/after the best residual correction.

Task	Subset	n	Rel. KL	Remove rel.	Acc. q4 $\rightarrow$ patch
PIQA	All	1838	0.940	2.398	0.801 $\rightarrow$ 0.804
PIQA	Teacher-correct	1490	0.938	2.527	0.974 $\rightarrow$ 0.977
PIQA	Boundary	39	0.941	2.090	0.000 $\rightarrow$ 0.154
ARC-Easy	All	570	0.961	2.418	0.770 $\rightarrow$ 0.775
ARC-Easy	Teacher-correct	450	0.970	2.425	0.960 $\rightarrow$ 0.969
ARC-Easy	Boundary	18	0.761	1.025	0.000 $\rightarrow$ 0.389

Table 3: Spectral residual-mode correction. Recovery is measured relative to the full oracle residual-correction gain.

Patch	Mean KL	Rel. KL	Gain recovery	$\|\Delta R_K\| \,/\, \|\Delta R\|$
q4 baseline	0.08459	1.000	0.0%	–
Full delta	0.07719	0.913	100.0%	1.000
Low-frequency $K = 64$	0.08261	0.977	26.8%	0.441
Low-frequency $K = 128$	0.08240	0.974	29.5%	0.508
High-frequency $K = 128$	0.08438	0.998	2.8%	0.238
Random $K = 128$	0.08367	0.989	12.3%	0.255

graph-frequency modes. We compare low-frequency modes with high-frequency modes and random orthonormal token-space subspaces of the same dimension.

Table 3 shows that low-frequency modes recover a nontrivial part of the oracle correction gain. With $K = 128$, low-frequency modes reduce KL from 0.08459 to 0.08240, recovering 29.5% of the full-delta gain. High-frequency modes recover only 2.8%, and random modes recover 12.3%. This is initial evidence that part of the behaviorally useful quantization error is not independent microscopic noise, but is organized as a smooth collective deformation on the teacher residual graph.

4 Implications and Limitations

These results suggest that low-bit compression can preserve the dominant layer-update drift while damaging a smaller residual component that still matters for output distributions and downstream choices. For fast model compression, this provides a diagnostic beyond aggregate output or reconstruction metrics: residual geometry can indicate where limited calibration, mixed precision, or correction capacity should be spent.

The proposed interventions are oracle probes. They use full-precision teacher residuals at the same input positions and are not deployable inference algorithms. The main behavioral results are from one model family, Llama-3.1-8B, one quantization backend, NF4 q4, and one late intermediate layer. We also find a trajectory-compatibility caveat: at layer 31, direct teacher-residual transplantation worsened KL (relative KL 1.23 at $\alpha = 0.3$, 2.36 at $\alpha = 0.75$), confirming that the useful object is residual structure expressed at compatible layers, not a raw teacher vector copied freely across trajectories.

The spectral experiment is also initial. Low-frequency modes recover more KL than high-frequency and random controls, but they also capture more residual-error norm. Stronger conclusions require larger token sets, energy-matched controls, multiple random seeds, and additional models. Nevertheless, the evidence supports a practical diagnostic claim: quantization damage is not purely uniform coordinate noise. It contains behaviorally relevant residual structure that aggregate metrics can miss.

Conclusion. We introduced a teacher-fixed residual-geometry diagnostic for low-bit language-model compression. On Llama-3.1-8B NF4 q4, structured residual correction reduces output KL, while removal, shuffling, and randomization degrade behavior. The same intervention improves teacher-faithful multiple-choice distributions, and low-frequency residual modes recover part of the oracle correction. These findings motivate residual-aware diagnostics for fast compression.

References

Ting-Yun Chang, Muru Zhang, Jesse Thomason, and Robin Jia. Why do some inputs break low-bit LLM quantization? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3410–3429, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.168. URL https://aclanthology.org/2025.emnlp-main.168/.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c3ba4962c05c49636d4c6206a97e9c8a-Paper-Conference.pdf.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=OUIFPHEgJU.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers, 2023. URL https://arxiv.org/abs/2210.17323.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In P. Gibbons, G. Pekhimenko, and C. De Sa, editors, *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38087–38099. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.