FROM HIDDEN STATES TO PROBABILITY MEASURES: MEASURING QUANTIZATION-INDUCED DEFORMATION IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention rows and next-token output distributions are probability measures. Prior work evaluates compressed models with accuracy, output KL, and answer flips; we extend this output-level evaluation to attention routing and treat both distributions as a measure-valued diagnostic layer. We compare discrepancies with and without ground geometry to characterize quantization-induced redistribution over positions and vocabulary outcomes. Using same-prefix, teacher-forced comparisons isolates local deformation from free-running history divergence. In the tested, comparatively mild INT8-versus-FP8 weights-only comparison, attention KL and W_1 distinguish all six model-domain pairs with a consistent direction; perplexity is significant in only $3/6$ and disagrees in direction even among significant pairs. Output KL has one directional exception. Under stronger INT4 weights-only quantization, perplexity also detects degradation. Adding activation quantization reverses the INT8/FP8 ordering in every tested pair, with model- and diagnostic-dependent magnitude and architecture-dependent layerwise shape. These diagnostics measure local redistribution and are not downstream-accuracy surrogates or full path-law comparisons.

1 INTRODUCTION

Perplexity records likelihood at the observed next token; hidden-state L^2 records coordinate differences. Neither locates how quantization redistributes attention or output probability mass. For a fixed query, an attention row is a probability measure over valid key positions; the output distribution is a probability measure over vocabulary outcomes. We compare a full-precision model M_{fp} with its quantized counterpart M_q at these measure-valued objects. The possibility of studying changes in Transformer behavior through probability-measure deformation motivates our focused comparison of attention rows and output distributions.

Residual and continuous-depth views motivate tracking hidden-state trajectories through a Transformer He et al. (2016); E (2017); Chen et al. (2018). But standard dot-product self-attention is not globally Lipschitz on unbounded inputs Kim et al. (2021), and available local bounds can depend on the input, sequence length, and attention concentration Castin et al. (2024); Yudin et al. (2025). Even when finite-context reachable states are bounded, composing local bounds across layers may be too coarse to assess a specific quantization perturbation. This concerns the usefulness of the bound, not a measured unusually large Lipschitz constant; Section 3 gives the technical argument.

We therefore compare attention rows over key positions and predictive distributions over the vocabulary. Under this view, quantization changes the mass assigned to positions and tokens. Deterministic greedy decoding from a fixed prompt is the degenerate single-trajectory case of a path law. Our experiments instead hold the ground-truth prefix $x_{\leq t}$ fixed for both models. This teacher-forced design separates local quantization-induced deformation from the history divergence of free-running generation.

KL measures discrepancies without ground geometry; Wasserstein-type diagnostics additionally use token-position or embedding geometry. We use them as representative, not canonical, choices. Together they test how quantization changes attention routing and next-token distributions, building on output-level fidelity evaluation rather than claiming a new divergence or quantizer.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our contributions are as follows.

- Our paired weights-only audit finds INT8 closer to full precision than FP8 by attention diagnostics in every tested model-domain pair; output KL has one directional exception. Perplexity's unstable direction in that comparison contrasts with its detection of degradation under stronger INT4 weights-only quantization.
- Comparing weights-only with weights-plus-activations, we document the INT8/FP8 reversal in every tested pair and examine its magnitude across models and diagnostics and its layerwise shape across architectures.
- Attention KL and W_1 distinguish routing anomalies from positional displacement; output KL and native-embedding transport show the limited extra severity resolution of that embedding geometry. Hidden-state L^2 is strongly confounded by depth and architecture.
- We relate same-prefix conditional KL to full autoregressive path-measure KL without treating dataset-prefix averages as path-KL estimators. Pooled downstream associations do not establish fine-grained weights-only or reliable item-level prediction.

2 Related Work

LLM quantization. Post-training LLM quantization includes weight-only, activation-aware, mixed-precision, and calibrated methods Dettmers et al. (2022); Frantar et al. (2023); Yao et al. (2022); Xiao et al. (2023); Lin et al. (2024); Shao et al. (2024). We use existing quantization settings to compare their effects on attention routing and next-token measures; we do not propose a quantizer.

Evaluation of language models. Perplexity measures observed-next-token likelihood, and accuracy summarizes task outcomes. Compressed models can retain similar accuracy while changing answers; KL and answer flips provide additional fidelity checks Dutta et al. (2024). These established limits motivate our comparison of attention routing.

Distribution matching in compression. Predictive-distribution matching is established in distillation and quantized compression Hinton et al. (2015); Polino et al. (2018). Across broad quantization cohorts, KLD and perplexity can correlate with downstream quality while losing ranking signal near the high-fidelity baseline; KLD measures disagreement volume more directly than the direction of disagreement Nikolić et al. (2026). We extend these output-level comparisons to attention routing distributions and compare attention KL with W_1 and output KL with D_{Sink} under their different ground geometries.

Dynamical and geometric views of neural networks. Residual networks He et al. (2016) motivate continuous-depth interpretations E (2017); Chen et al. (2018), in which Lipschitz bounds control error propagation. Standard dot-product self-attention is not globally Lipschitz on unbounded input domains Kim et al. (2021); available local bounds can depend on the input, sequence length, and attention concentration Castin et al. (2024); Yudin et al. (2025). Section 3 develops the consequences for composing layer-wise bounds and comparing probability measures; Appendix A.1 gives the standard discrete bound.

Optimal transport and Sinkhorn distances. Optimal transport provides a geometry-aware way to compare probability measures Villani (2009); Peyre & Cuturi (2019). Entropic regularization and Sinkhorn iterations make transport discrepancies computationally tractable in machine-learning settings Cuturi (2013); Sinkhorn divergences further introduce a debiased geometry-aware discrepancy Feydy et al. (2019). Recent work has also used Wasserstein distances as diagnostics for internal structure in compressed LLMs. Sawmya et al. (2025) identify "Wasserstein neurons" by comparing normalized neuron-output distributions to a Gaussian using one-dimensional Wasserstein distances, and relate these neurons to sparsification sensitivity. Our use of Wasserstein distances is complementary: we compare probability measures induced by full-precision and quantized models, such as attention rows and next-token distributions, rather than measuring the non-Gaussianity of individual neuron outputs.

We apply transport to the induced measures, without assuming that Transformer computation solves an optimal transport problem.

3 METHODOLOGY

3.1 COMPARISON OBJECTS AND CONDITIONING

We treat attention rows over token positions and the output layer's distribution over the vocabulary as the concrete measures whose quantization-induced deformation we measure, and we select KL divergence and Wasserstein-type distances as representative, tractable ways to read off that deformation. Other divergences, ground geometries, or path-space functionals may be more appropriate for specific diagnostic goals.

Residual architectures have been interpreted as discretizations of continuous dynamical systems He et al. (2016); E (2017); Chen et al. (2018), and under this reading, a natural way to bound how far a quantized model's hidden state can drift from the full-precision one is to bound the Lipschitz constant of each layer and compose these bounds across depth, exactly as one bounds discretization error and perturbation propagation for an ODE solver via a Lipschitz bound on the vector field. For layers built from self-attention, this program meets a specific difficulty. The map itself composes the bilinear map $(Q, K) \mapsto QK^\top$, with Q and K themselves depending linearly on the input, with a softmax, and available bounds on the resulting sensitivity are input-dependent: on an unbounded input domain, standard dot-product self-attention is not globally Lipschitz Kim et al. (2021). This unbounded-domain fact alone does not settle the matter for a fixed model on finite-context inputs, whose reachable states occupy some bounded region; the argument that matters in practice is instead about local bounds on that bounded region. Even there, available local Lipschitz bounds can depend on the input, on the sequence length, and on how concentrated the resulting attention distribution is, rather than being single fixed properties of the layer Castin et al. (2024); Yudin et al. (2025). A bound compiled by composing such input-dependent, per-layer bounds across many layers can therefore become too coarse to certify anything useful about the specific perturbation that a given quantization scheme introduces, even though no single number in the chain need be pathological on its own. Our concern is the informativeness of the composed bound for a particular perturbation; we have not measured an unusually large Lipschitz constant for softmax or self-attention.

Pointwise Euclidean hidden-state distances inherit this difficulty in a specific way: by construction, they measure coordinate discrepancy after whatever redistribution of probability mass attention has performed, without directly measuring that redistribution, so they conflate quantization-induced change with architecture-dependent depth effects that are not attributable solely to quantization severity. Section 4 tests this empirically, showing that hidden-state L^2 is strongly depth-dependent, with a signature specific to each architecture, without treating hidden-state divergence as a degradation score.

We therefore analyze the probability measures attention and the output layer already induce. For a fixed query position, the attention row is a distribution over the valid key positions that determines how information from other positions is mixed into the current one; the output layer's logits, after softmax, are a distribution over the vocabulary that determines which continuation is realized. We ask how quantization redistributes mass within these two distributions. Distribution matching is not new in model compression Hinton et al. (2015); Polino et al. (2018); the contribution here is to treat these two measure-valued objects as a joint diagnostic layer for quantization, and to instantiate that layer with both a ground-geometry-free discrepancy (KL) and geometry-aware ones (Wasserstein-type distances), comparing what each reveals.

Let M_{fp} denote a full-precision language model and M_q its quantized counterpart. Given a prefix $x_{\leq t}$, both models define next-token distributions over the vocabulary $\mathcal{V}$:

$$p_{\mathrm{fp}}(\cdot \mid x_{\leq t}), \qquad p_q(\cdot \mid x_{\leq t}).$$

These are two probability measures on the same vocabulary $\mathcal{V}$ – equivalently, two points of the simplex $\Delta^{|\mathcal{V}|-1}$ – and a discrepancy between them can be computed regardless of what each is conditioned on. Interpreting that discrepancy as the local deformation induced by quantization at a single conditioning event requires holding the conditioning context fixed across the two models. This is not automatic: if M_{fp} and M_q are run as free-running autoregressive generators and sampled

token-by-token, the sequences they produce generally diverge as soon as one sampled token differs, after which $x^{\text{fp}}_{<t} \neq x^{q}_{<t}$ and every subsequent conditional distribution being compared is conditioned on a different history. A time-wise distance between such trajectories then mixes two effects that we want to keep separate: the local deformation that quantization induces at a given conditioning event, and the accumulated consequence of the two models having taken different paths to get there.

We isolate the first effect by using a teacher-forced, ground-truth-prefix evaluation: for a dataset $\mathcal{D} = \{x^{(i)}_{1:T_i}\}_{i=1}^{N}$, we condition both M_{fp} and M_q on the same dataset prefix $x^{(i)}_{\leq t}$ at every valid position, with neither model using its own sampled continuation. Because the conditioning variable is held fixed across the two models, the discrepancy between $p_{\text{fp}}(\cdot \mid x_{\leq t})$ and $p_q(\cdot \mid x_{\leq t})$ is not contaminated by history divergence, and it isolates the local deformation induced by quantization at that step in a same-prefix conditional comparison. This is a fixed-prefix, single-step conditional evaluation: the model is not repeatedly conditioned on its own sampled outputs in the main experiments, and the resulting object is a local, dataset-prefix-averaged quantity, not an expectation taken along either model's own rollout distribution.

In our exploratory free-running comparison, direct token-embedding trajectory distance did not behave as a stable diagnostic of semantic generation difference; see Appendix A.2.

3.2 Diagnostics

Perplexity measures how much probability a model assigns to the observed next token. It does not directly measure how probability mass is rearranged among alternative tokens.

Output KL is a directed, ground-geometry-free discrepancy from the full-precision to the quantized next-token distribution. It asks how much predictive mass the full-precision model places on outcomes that the quantized model suppresses or reallocates on the vocabulary simplex. KL is highly sensitive to probability mass that is assigned by the full-precision model but suppressed by the quantized model. This makes it useful as a distributional collapse detector and as a strong practical output-level diagnostic, but it can also take very large values when distributions nearly mismatch in support.

Where output KL is ground-geometry-free, the embedding-space D_{Sink} diagnostic asks how far predictive mass moves on a chosen token geometry: it distinguishes mass that shifts to a nearby token from mass that shifts to a distant one, relative to that geometry. We use the full-precision model's input embedding table as the static ground geometry and truncate support to the union of each model's top-64 tokens. This optional geometry-aware diagnostic is conditioned on that particular native geometry rather than a canonical or task-calibrated one. It is derived from entropically regularized transport but is not a canonical W_1, W_2, or exact full-vocabulary Sinkhorn divergence.

For a fixed query position, attention KL asks how sharply the full-precision routing distribution over key positions has been deformed, and in particular how much routing probability the full-precision model assigns to a position that the quantized model suppresses; like output KL, it is ground-geometry-free and correspondingly sensitive to near-support mismatch and localized anomalies. Attention W_1 uses the natural one-dimensional ordered geometry of token positions to measure the positional displacement of attention mass. Its closed-form cumulative computation is a convenience of this geometry, not a claim that W_1 is uniquely correct for measuring attention displacement. Definitions, corpus-level aggregation, support and cost construction, and implementation details are in Appendix A; document-level aggregation is specified in Appendix B.

3.3 Connection to Path Measures

Although our experiments evaluate fixed-prefix single-step conditional distributions, conditional KL terms are not merely heuristic local probes. For finite-length autoregressive models, the KL divergence between path measures decomposes exactly into conditional next-token KL terms. Let $P^{(T)}_{\text{fp}}$ and $P^{(T)}_{q}$ denote the length-T sequence distributions induced by M_{fp} and M_q. In this decomposition, the outer expectation over prefixes is taken under the full-precision model's own rollout distribution, $X_{<t} \sim P_{\text{fp}}$; Appendix A gives the product-form definitions and chain-rule derivation.

Thus single-step conditional KL is an exact component of sequence-level, path-measure KL only under that rollout expectation. This is a distinct object from what our main experiments actually compute: Section 4 averages the same conditional KL over ground-truth dataset prefixes $x_{\leq t}^{(i)}$, not over prefixes sampled from M_{fp}'s rollout. The two prefix distributions need not coincide, so the reported document-level statistic should be read as a local, dataset-conditioned diagnostic that is mathematically related to a component of path-measure KL, and not asserted to be an exact estimator of $D_{\mathrm{KL}}(P_{\mathrm{fp}}^{(T)} \| P_q^{(T)})$ itself. Even granting the rollout-expectation reading, the equality is asymmetric: it probes prefixes typical under M_{fp}, but not necessarily prefixes that become likely only under M_q's own errors. The directionality of path-measure KL can be addressed with a reverse or symmetric comparison; causal or adapted transport addresses a distinct issue by constraining couplings to respect temporal information flow, which we return to in Section 6.

4 Experiments

4.1 Experimental Setup

We evaluate three model families, Qwen3-4B, Mistral-7B, and Llama-3.2-3B, over two datasets: IFBench (300 documents) and WikiText (2,891 documents). For each model, we compare five quantization settings: INT8 weights-only, FP8 weights-only, INT4 weights-only, INT8 weights-plus-activations, and FP8 weights-plus-activations. Quantization is performed with the `optimum-quanto` library using its default data-free linear quantization, with no calibration dataset; installed library versions are recorded automatically at quantization time. This yields $3 \times 5 \times 2 = 30$ model-setting-dataset cells (equivalently, $3 \times 5 = 15$ quantized model configurations pooled across the two datasets). Broader corpus coverage was limited by available computational resources; C4 and FineWeb were therefore excluded from the main analysis rather than evaluated with substantially smaller and uneven sample sizes.

We report perplexity ratios, output KL, output embedding-space Sinkhorn diagnostics, attention KL, attention W_1, and hidden-state L^2. Unless otherwise stated, comparisons are performed at the document level. The mild INT8-versus-FP8 weights-only audit (Section 4) is aggregated across the $3 \times 2 = 6$ model-domain pairs; the INT4 audit uses the same six pairs against each of INT8 and FP8; and the downstream-correlation analysis (Section 4.7) is aggregated across all 15 model-setting configurations, pooling documents from both datasets.

4.2 Statistical Testing Procedure

Every diagnostic is reduced to one summary value per document, and FP8 and INT8 are compared on identical documents. For each of the five diagnostics (perplexity ratio, output KL, output D_{Sink}, attention KL, and attention W_1), we apply a two-sided Wilcoxon signed-rank test to paired differences (FP8 minus INT8) and report matched-pairs rank-biserial correlation r. Holm–Bonferroni correction is applied separately within each diagnostic's family of six model-domain tests; statistical significance means $p_{\mathrm{Holm}} < 0.05$. Appendix B contains the exact implementation and the fully specified per-cell results that are available; per-cell effect sizes for output KL, attention KL, attention W_1, and Sinkhorn are reported there as aggregate ranges rather than a full per-cell table, as noted in the appendix.

4.3 The Insufficiency and Directional Instability of Perplexity in the Comparatively Mild INT8-versus-FP8 Weights-Only Comparison

We compare INT8 and FP8 weights-only quantization using paired document-level tests (Section 4.2). Both formats often remain close to full precision in these tested pairs.

Across the 6 model-domain pairs, perplexity is significant after Holm correction in only $3/6$. Its direction differs even among significant pairs: Llama-3.2-3B on WikiText has $r = +0.529$ (FP8 deviates more), whereas Mistral-7B on WikiText has $r = -0.314$ (INT8 deviates more). Perplexity sometimes favors a quantized model over full precision despite a nonzero distributional perturbation.

Attention KL and attention W_1 are significant in all $6/6$ pairs with uniform direction. Output KL is also significant in $6/6$, but Qwen3-4B on WikiText reverses the direction; sensitivity to extremely

short inputs strongly amplifies, but does not fully explain, that exception. FP8 deviates more than INT8 in the other five pairs. Table 1 reports effect-size ranges and the paired-test results.

Table 1: Document-level discriminability of INT8 versus FP8 under weights-only quantization across 6 model-domain pairs (three models $\times$ two datasets). Tests are paired Wilcoxon signed-rank tests with matched-pairs rank-biserial effect size (Appendix B); Holm correction is applied within each diagnostic's family of six tests. The counts establish discrimination under the stated aggregation protocol, not an aggregation-independent ranking of intrinsic diagnostic quality.

Diagnostic	**Sig. (Holm)**	**Direction**	**Effect size** $\lvert r \rvert$
Perplexity ratio	$3/6$	$5:1$, and inconsistent even when significant*	0.006–0.529
Output KL	$6/6$	$5/6$ FP8 $>$ INT8**	0.20–1.00
Output D_{Sink}	$6/6$	$6/6$ FP8 $>$ INT8	0.75–1.00
Attention KL	$6/6$	$6/6$ FP8 $>$ INT8	0.97–1.00
Attention W_1	$6/6$	$6/6$ FP8 $>$ INT8	0.996–1.00

*Llama-3.2-3B on both datasets and Mistral-7B on WikiText are individually significant, but Llama's direction (FP8>INT8) is the opposite of Mistral's (INT8>FP8); see the discussion below. **Qwen3-4B/WikiText shows the reverse direction (INT8 $>$ FP8), strongly amplified by a model-specific sensitivity to extremely short inputs (see the discussion above). Output D_{Sink} does not share this exception.

Perplexity records observed-token likelihood; attention KL and W_1 detect routing changes, and output KL detects predictive-mass reallocation. Our paired audit compares established perplexity and output-KL diagnostics with attention-routing measures under the same quantization settings.

Across the three models and IFBench/WikiText, we also pair documents for INT4-WO against each of INT8-WO and FP8-WO (12 contrasts per diagnostic). Two-sided Wilcoxon signed-rank tests with Holm correction find higher INT4 perplexity ratios in (10/12) contrasts; both nonsignificant contrasts are Mistral-7B/WikiText (versus FP8: raw (p=0.7544); versus INT8: raw (p=0.6408)). Output KL, output D_{Sink}, attention KL, and attention W_1 show greater INT4 deformation in all (48/48) contrasts after Holm correction (matched-pairs rank-biserial $|r| \geq 0.999$).

4.4 A Qualitative Inversion Between Weights-Only and Activation Quantization

The ordering between INT8 and FP8 changes with the quantization target. For the tested INT8-versus-FP8 weights-only pairs above, INT8 is closer to full precision in all $6/6$ pairs by attention diagnostics and $5/6$ by output KL; Qwen3-4B on WikiText is the exception. FP8's limited mantissa could produce weight perturbations amplified by attention and output softmax operations, but we have not tested this mechanism directly.

With activation quantization, INT8 deviates more than FP8 in every tested pair. The reversal is consistent in direction across the three architectures, while its magnitude depends on model and diagnostic: the WikiText document-mean INT8/FP8 deviation ratio ranges from $1.58\times$ for Mistral-7B output KL to $345.75\times$ for Llama-3.2-3B attention KL (Table 2). No single multiplicative factor describes the reversal.

The INT8-plus-activations deformation persists across layers in all three architectures, with different onset shapes; FP8 peaks cleanly at the terminal layer only for Mistral-7B (Section 5, Table 3). Activation outliers and INT8's limited dynamic range are a plausible explanation consistent with prior work on mixed-precision remedies Dettmers et al. (2022), not a mechanism verified here by activation histograms or saturation counts.

Our contribution is diagnostic rather than causal or algorithmic: attention and output measures reveal differences across quantization targets and layers that endpoint scalars do not resolve. The INT8/FP8 ordering therefore depends on what is quantized, as well as on precision format.

4.5 KL and Wasserstein Have Distinct Diagnostic Roles

Attention KL detects localized near-support routing mismatches; attention W_1 measures positional mass displacement. Attention W_1 has a closed-form cumulative computation, strong binary discrim-

Table 2: Ratio of INT8's mean deviation from full precision to FP8's mean deviation, under weights-plus-activations quantization, document-mean values on WikiText. Values above 1 indicate INT8 deviates more than FP8 for that model and diagnostic; the direction is consistent across the tested models, but the magnitude spans more than two orders of magnitude.

Model	**PPL ratio**	**Output KL**	**Attention KL**	**Attention W_1**	**Output D_{Sink}**
Qwen3-4B	140,366	23.29	52.82	5.98	2.43
Mistral-7B	25.08	1.58	48.997	9.90	1.14
Llama-3.2-3B	34,981	184.85	345.75	20.11	3.84

ination, and smoother layerwise profiles than KL in the tested activation-quantized configurations (Table 3). The smoothness is empirical, not a stability theorem; KL remains sensitive to near-support anomalies.

At the output, D_{Sink} uses the chosen native embedding geometry, while KL is ground-geometry-free. Both are significant in all $6/6$ tested weights-only INT8-versus-FP8 pairs, but D_{Sink} has uniform direction and output KL has one directional exception (Table 1). Output KL offers stronger continuous severity resolution and is our practical default; D_{Sink} is an optional geometry-aware probe. The diagnostics have different uses, without a universal ranking.

4.6 The Output Sinkhorn Diagnostic Provides Limited Incremental Value

Output D_{Sink} distinguishes nearby from distant shifts in predictive mass under the chosen native embedding geometry. In the tested settings, it detects broad perturbations also seen by output KL, with limited incremental value relative to its iterative cost. The finding is conditional on the native input embeddings, which need not encode the distinctions most relevant to quantization; it does not rule out other transport geometries. Context-dependent, output-head-based, or task-calibrated geometries could change the diagnostic value of output transport.

4.7 Relationship to Downstream Task Accuracy

The diagnostics measure fixed-prefix, teacher-forced deformation of attention and next-token probability measures, not downstream accuracy itself. Across the pooled fifteen configurations, all ten configuration-level Spearman correlations between the five diagnostics and two downstream outcomes remain significant after Holm correction. A substantial part of this pooled association reflects the severity gap between mild weights-only settings and severe activation quantization, with INT4 weights-only between them. Restricted to the nine weights-only configurations, none of the ten correlations remains significant after Holm correction, so this does not establish a fine-grained relationship. At item level, $5/90$ one-sided and $4/90$ two-sided tests remain significant after Holm correction; every significant test is confined to INT8-plus-activations and HellaSwag, across Qwen3-4B, Mistral-7B, and Llama-3.2-3B. These results suggest a pooled configuration-level association with downstream behavior, but do not establish reliable item-level prediction or a surrogate for downstream accuracy. The full analysis and Tables A2–A4 are in Appendix C.

5 Layer-wise Diagnostic Behavior

5.1 Attention KL as an Anomaly Detector and W_1 as a Continuous Gauge

Table 3 summarizes attention KL and W_1 by layer under weights-plus-activations quantization on WikiText. INT8 deformation is large and persistent across all three architectures, with an early KL jump in Mistral-7B and Llama-3.2-3B but already large first-layer KL in Qwen3-4B. With FP8, both diagnostics peak at the terminal layer only for Mistral-7B; Qwen3-4B and Llama-3.2-3B peak earlier. The onset and concentration of deformation depend on architecture.

Table 3: Layerwise deformation summary under weights-plus-activations quantization, WikiText, document-mean values normalized by token count before averaging across documents. "Peak layer" is the layer index (of the total layer count for that model) at which the per-layer statistic is largest; "final/preceding ratio" is the last layer's value divided by the second-to-last layer's.

Model	Config, diagnostic	Peak layer	Final/preceding ratio
Mistral-7B	fp8+act, attn. KL	32/32 (terminal)	$4.78\times$
Mistral-7B	fp8+act, attn. W_1	32/32 (terminal)	$1.88\times$
Qwen3-4B	fp8+act, attn. KL	30/36	$0.45\times$ (a decrease)
Qwen3-4B	fp8+act, attn. W_1	17/36	$0.75\times$ (a decrease)
Llama-3.2-3B	fp8+act, attn. KL	14/28 (mid-network)	$0.78\times$
Llama-3.2-3B	fp8+act, attn. W_1	14/28 (mid-network)	$1.04\times$
Mistral-7B	int8+act, attn. KL (layer 1→2 jump)	–	$23.5\times$
Llama-3.2-3B	int8+act, attn. KL (layer 1→2 jump)	–	$4.4\times$
Qwen3-4B	int8+act, attn. KL	34/36	already large from layer 1, no jump

5.2 Cross-Architecture Check: Three Model Families

Across the three tested architectures, the persistent INT8-plus-activations deformation has different onset shapes, and FP8-plus-activations terminal-layer concentration holds cleanly only for Mistral-7B. The layerwise pattern is not precision-format-universal (Table 3).

5.3 Hidden-State L^2 Is Strongly Depth-Dependent

Hidden-state L^2 records coordinate discrepancy after attention redistribution rather than directly measuring attention or output measures. It is very strongly correlated with layer index in most tested settings; Qwen3-4B under INT8-plus-activations is an exception. The final-layer discontinuity differs in sign and magnitude across model families, while its within-model pattern persists across weights-only precision levels. These depth- and architecture-dependent effects support the concern in Section 3. Hidden-state L^2 is useful for observing trajectory divergence but unreliable as a primary quantization-severity diagnostic (Appendix D, Table A5).

6 Limitations and Future Work

Our experiments measure local, same-prefix conditional deformation, not full free-running path laws. Dataset-prefix averages are not estimators of full path-measure KL (Section 3). Ordinary optimal transport on sequence space can use path-dependent costs, but does not by itself enforce nonanticipative couplings that respect the autoregressive filtration. Nested distances and causal or bicausal transport are natural candidates for future comparison; we have not implemented or evaluated them (Appendix E).

The output transport diagnostic depends on the chosen native embedding geometry, which is not canonical or task-calibrated; its limited incremental value is conditional on that choice (Appendix E). These diagnostics are not established as fine-grained or item-level downstream predictors or as surrogates for accuracy (Section 4.7). Inferential tests use documents as units of observation. Coverage is limited to three model families, two datasets, and five quantization settings: INT8, FP8, and INT4 weights-only, plus INT8 and FP8 weights-plus-activations. These experiments use naive quantization settings; whether the observed INT8/FP8 ordering persists under calibrated methods such as SmoothQuant and AWQ remains untested (Appendix E).

7 Conclusion

Our results provide empirical evidence that comparing attention and output probability-measure deformation at matched prefixes is useful for diagnosing quantization. In the tested, comparatively mild INT8-versus-FP8 weights-only comparison, attention-level diagnostics consistently find INT8 closer to full precision than FP8 while perplexity is directionally unreliable; output KL has one directional exception. Under stronger INT4 weights-only quantization, perplexity also detects degradation. Adding activation quantization reverses the INT8/FP8 ordering in every tested pair, with

model- and diagnostic-dependent magnitude and architecture-dependent layerwise shape. Hidden-state L^2 is confounded by depth and architecture, while the output transport diagnostic depends on the chosen native embedding geometry. The same-prefix, dataset-conditioned diagnostics measure local deformation and are not downstream-accuracy substitutes: pooled configuration-level associations do not establish a fine-grained relationship within the nine weights-only configurations after Holm correction or reliable item-level prediction. Dataset-prefix conditional KL is related by the chain rule to full autoregressive path-measure KL, whose outer expectation uses full-precision rollout prefixes; the dataset-prefix average is not a path-KL estimator. Comparison of full path measures using transport formulations that respect temporal information structure is a natural next stage.

Reproducibility Statement

The three model families, two datasets, and five quantization settings used in our experiments, together with the resulting $3 \times 5 \times 2 = 30$ model-setting-dataset cells and the quantization library and calibration procedure, are fully specified in Section 4.1. The statistical testing procedure — the paired Wilcoxon signed-rank test, the matched-pairs rank-biserial effect size, and the Holm–Bonferroni correction — is described in Section 4.2 and given in full detail, including the pairing-integrity verification, in Appendix B. Exact definitions of all five diagnostics, including all hyperparameters used in the implementation (e.g., $\varepsilon = 0.05$, top-64 support truncation, $I = 300$ Sinkhorn iterations, float64 precision), are given in Appendix A. The full downstream analysis methodology and the hidden-state L^2 computation are given in Appendices C and D, respectively.

AI Use Statement

In this work, we used generative AI tools to interpret results, to create or edit software code, and to edit the manuscript to improve readability. Specifically, generative AI tools were used (1) to independently recompute and cross-check statistical results reported in this paper — including significance tests, effect sizes, and correlation coefficients — against the underlying per-document data, (2) to debug the experiment code, with all resulting fixes reviewed and verified by the authors, and (3) to proofread the manuscript and revise its English into natural, idiomatic phrasing that reads clearly as an academic paper. All AI-assisted verifications, code changes, and edits were reviewed by the authors against the source data and the authors' original intent before inclusion. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Julio Backhoff, Mathias Beiglbock, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017.

Valerie Castin, Pierre Ablin, and Gabriel Peyre. How smooth is attention? In *International Conference on Machine Learning*, 2024.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, 2022.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e0e956681b04ac126679e8c7dd706b2e-Abstract-Conference.html.

Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

Jean Feydy, Thibault Sejourne, Francois-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyre. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *International Conference on Machine Learning*, 2021.

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 *MLSys*, 2024.

Miloš Nikolić, Ali Hadi Zadeh, Enrique Torres Sanchez, and Andreas Moshovos. Displacement is not direction: Evaluating fidelity metrics for quantized LLM deployment, 2026. URL `https://arxiv.org/abs/2606.19558`.

Gabriel Peyre and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Georg Ch. Pflug and Alois Pichler. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 22(1):1–23, 2012.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. In *International Conference on Learning Representations*, 2018.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhauser, Cham, 2015.

Shashata Sawmya, Linghao Kong, Ilia Markov, Dan Alistarh, and Nir N. Shavit. Wasserstein distances, neuronal entanglement, and sparsity. In *International Conference on Learning Representations*, 2025.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024.

Cedric Villani. *Optimal Transport: Old and New*. Springer, 2009.

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 *International Conference on Machine Learning*, 2023.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*, 2022.

Nikolay Yudin, Alexander Gaponov, Sergei Kudriashov, and Maxim Rakhuba. Pay attention to attention distribution: A new local Lipschitz bound for transformers. *arXiv preprint arXiv:2507.07814*, 2025.

A Additional Methodological Details

A.1 Discrete Grönwall Illustration

The standard form of this bound is the discrete Grönwall inequality. For an ODE $\dot{h}(t) = F(t, h(t))$ with Euler updates $h_{k+1} = h_k + \Delta t\, F(t_k, h_k)$, if F is L-Lipschitz in h and the local truncation error is bounded by $C\Delta t^2$, the global error e_k after k steps satisfies $e_{k+1} \le (1 + L\Delta t)e_k + C\Delta t^2$, so that after $n = T/\Delta t$ steps,

$$e_n \;\le\; C\Delta t^2 \sum_{j=0}^{n-1} (1 + L\Delta t)^j \;=\; \frac{C\Delta t}{L}\left((1 + L\Delta t)^n - 1\right) \;\le\; \frac{C\Delta t}{L}\left(e^{LT} - 1\right),$$

with the $L \to 0$ limit giving the linear bound $CT\Delta t$. We do not claim that quantized Transformer computation obeys this bound in the form stated; the derivation is included only to make explicit, in a standard setting, why the Lipschitz constant enters as an error-amplification factor across steps, which is the property that the argument below relies on.

A.2 Free-Running Token-Embedding Exploration

We also explored comparing free-running generated sequences directly, using distances between token embeddings along the two trajectories. Once histories diverge, the conditional distributions being compared are conditioned on different histories, so such a comparison mixes the local deformation quantization induces with the accumulated consequence of the two models having diverged, which is why we use a same-prefix design in the main experiments instead. Separately, and as an empirical observation rather than a claim we derive from the point above, this direct token-embedding trajectory distance did not behave as a stable diagnostic of semantic generation difference in our exploration. We treat this as a methodological observation motivating the shared-prefix design used here, not as evidence that free-running comparison is uninformative in general: principled path-space estimators or diagnostics built from collections of trajectories remain possible and are discussed as future work in Section 6.

A.3 Perplexity and Output KL Definitions

For a model M with conditional probabilities p_M, the corpus-level perplexity is

$$\mathrm{PPL}(M; \mathcal{D}) = \exp\left(-\frac{1}{\sum_i (T_i - 1)} \sum_{i=1}^{N} \sum_{t=1}^{T_i - 1} \log p_M(x_{t+1}^{(i)} \mid x_{\le t}^{(i)})\right).$$

At each prefix $x_{\le t}$, we compute the directed KL divergence Cover & Thomas (2006)

$$D_{\mathrm{KL}}(p_{\mathrm{fp}}(\cdot \mid x_{\le t}) \,\|\, p_q(\cdot \mid x_{\le t})) = \sum_{v \in \mathcal{V}} p_{\mathrm{fp}}(v \mid x_{\le t}) \log \frac{p_{\mathrm{fp}}(v \mid x_{\le t})}{p_q(v \mid x_{\le t})}.$$

The corpus-level output KL is the token-average

$$\overline{D_{\mathrm{KL}\mathrm{out}}} = \frac{1}{\sum_i (T_i - 1)} \sum_{i=1}^{N} \sum_{t=1}^{T_i - 1} D_{\mathrm{KL}}\Big(p_{\mathrm{fp}}(\cdot \mid x_{\le t}^{(i)}) \,\|\, p_q(\cdot \mid x_{\le t}^{(i)})\Big).$$

A.4 Truncated-Support Entropic-Transport Diagnostic

To incorporate token geometry, we compare output distributions on a truncated vocabulary support using the full-precision model's input embedding table. Let $e_v \in \mathbb{R}^d$ denote the unnormalized input embedding of token v. At each prefix, we take the union of the top $k = 64$ tokens under the full-precision and quantized logits,

$$S_t = \mathrm{TopK}_{64}(p_{\mathrm{fp}}(\cdot \mid x_{\le t})) \cup \mathrm{TopK}_{64}(p_q(\cdot \mid x_{\le t})),$$

so $|S_t| \le 128$, and renormalize each distribution on S_t. This produces conditional distributions on the selected support; it is not the exact full-vocabulary transport problem because tail mass outside S_t is discarded.

For $u, v \in S_t$, the implementation uses the Euclidean ground cost

$$C_{uv} = \|e_u - e_v\|_2.$$

At each token position, the implementation selects the union of the top 64 logits from each model, yielding a support of at most 128 unique tokens. Restricted probabilities are obtained by applying softmax only to the selected logits, which is equivalent to conditioning the original softmax distribution on that support. Consequently, D_{Sink} ignores how much total probability each model assigns outside the selected union and is not guaranteed to approximate the full-vocabulary transport value without an additional tail-mass bound. We use it as a controlled, computationally tractable geometry-aware diagnostic rather than as an exact full-support distance.

The full-precision input embedding table supplies the unnormalized vectors used in the Euclidean cost matrix. For support size $m \leq 128$, cost-matrix construction is $O(m^2 d)$. Each log-domain matrix-scaling iteration is $O(m^2)$; debiasing requires three solves, each with $I = 300$ fixed iterations. We did not benchmark wall-clock runtime in this study, so statements about cost refer only to this additional algorithmic work relative to non-iterative KL computation.

Transport component and implementation. Let $\pi_\varepsilon(p, q)$ be the entropically regularized coupling obtained by minimizing

$$\langle \pi, C \rangle + \varepsilon \sum_{u,v \in S_t} \pi_{uv}(\log \pi_{uv} - 1) \quad \text{over } \pi \in \Pi(p, q),$$

with $\varepsilon = 0.05$, following the entropic-regularization framework of Cuturi Cuturi (2013). The code then records the transport part of that regularized plan,

$$T_\varepsilon(p, q) = \langle \pi_\varepsilon(p, q), C \rangle,$$

and forms the debiased quantity

$$S_\varepsilon^{\mathrm{cost}}(p, q) = T_\varepsilon(p, q) - \frac{1}{2} T_\varepsilon(p, p) - \frac{1}{2} T_\varepsilon(q, q).$$

The reported statistic is

$$D_{\mathrm{Sink}}(p, q) = \sqrt{\max\{S_\varepsilon^{\mathrm{cost}}(p, q), 0\}}.$$

The three-term subtraction mirrors the debiasing structure of Sinkhorn divergences Feydy et al. (2019), but here it is applied to the transport components of regularized plans rather than to the full regularized objectives. All three transport solves are performed in float64 with 300 fixed log-domain Sinkhorn iterations.

With $m = |S_t| \leq 128$ and $I = 300$, each of the three matrix-scaling solves requires $O(Im^2)$ updates after constructing the $m \times m$ cost matrix, whereas KL requires no iterative transport solve. We assess whether this geometry adds diagnostic value given the additional algorithmic work.

The diagnostic supplies geometry that KL ignores: it distinguishes mass moved between nearby embeddings from mass moved between distant embeddings. Its cost is additional algorithmic overhead, not an asserted deployment bottleneck.

A.5 Attention KL and Positional Wasserstein Details

For layer ℓ, head h, and query position t, let

$$a_{\mathrm{fp},t}^{\ell,h}, \quad a_{q,t}^{\ell,h} \in \Delta^{L-1}$$

denote the attention rows of the full-precision and quantized models over the L valid key positions. We compute attention KL as

$$D_{\mathrm{KL}}(a_{\mathrm{fp},t}^{\ell,h} \| a_{q,t}^{\ell,h}) = \sum_{s=1}^{L} a_{\mathrm{fp},t}^{\ell,h}(s) \log \frac{a_{\mathrm{fp},t}^{\ell,h}(s)}{a_{q,t}^{\ell,h}(s)}.$$

A.6 One-Dimensional W_1 Identity

Since attention positions lie on a one-dimensional ordered grid with unit spacing, the Wasserstein-1 distance has the cumulative-sum identity Villani (2009); Santambrogio (2015). This directly measures how far mass has been rerouted along the position geometry, not merely whether it moved:

$$W_1(a, b) = \sum_{s=1}^{L-1} \left| \sum_{r=1}^{s} (a_r - b_r) \right|.$$

The implementation reports the length-normalized row statistic $W_1(a, b)/L$, then aggregates it over heads, valid query positions, layers, and finally documents as specified in Appendix B. Unlike the output Sinkhorn diagnostic, attention W_1 requires only cumulative sums and no iterative optimization; this closed form is a computational convenience of the position geometry, not a claim that W_1 is uniquely correct for measuring attention displacement. Its geometry is also operationally interpretable: moving attention mass from a nearby token to a distant token changes the transport cost according to positional displacement. The following gives a short derivation of the identity.

For probability vectors a, b on the ordered unit grid $\{1, \ldots, L\}$, let $F_a(s) = \sum_{r \le s} a_r$ and $F_b(s) = \sum_{r \le s} b_r$. Any transport plan must move the net excess mass $F_a(s) - F_b(s)$ across the cut between positions s and $s + 1$. Summing the absolute mass crossing each unit-length cut gives

$$W_1(a, b) = \sum_{s=1}^{L-1} |F_a(s) - F_b(s)|,$$

which is the discrete specialization of the standard one-dimensional identity $W_1(\mu, \nu) = \int |F_\mu - F_\nu|$ Villani (2009); Santambrogio (2015). The reported row statistic divides this value by L to reduce the direct scale dependence on sequence length.

A.7 KL Chain Rule for Autoregressive Paths

Let $P_{\text{fp}}^{(T)}$ and $P_q^{(T)}$ be the length-T sequence distributions induced by M_{fp} and M_q:

$$P_{\text{fp}}^{(T)}(x_{1:T}) = \prod_{t=1}^{T} p_{\text{fp}}(x_t \mid x_{<t}), \qquad P_q^{(T)}(x_{1:T}) = \prod_{t=1}^{T} p_q(x_t \mid x_{<t}).$$

Then the chain rule for KL gives Cover & Thomas (2006)

$$D_{\text{KL}}\left(P_{\text{fp}}^{(T)} \| P_q^{(T)}\right) = \sum_{t=1}^{T} \mathbb{E}_{X_{<t} \sim P_{\text{fp}}} \left[D_{\text{KL}}(p_{\text{fp}}(\cdot \mid X_{<t}) \| p_q(\cdot \mid X_{<t})) \right].$$

Using the autoregressive factorizations above,

$$\log \frac{P_{\text{fp}}^{(T)}(X_{1:T})}{P_q^{(T)}(X_{1:T})} = \sum_{t=1}^{T} \log \frac{p_{\text{fp}}(X_t \mid X_{<t})}{p_q(X_t \mid X_{<t})}.$$

Taking expectation under $P_{\text{fp}}^{(T)}$, interchanging the finite sum and expectation, and conditioning on $X_{<t}$ yields the chain-rule expression above Cover & Thomas (2006). This equality is exact, but asymmetric because all prefixes are averaged under the full-precision path measure.

B Statistical Methods and Full Paired Test Results

Design. For each of the 6 model-domain pairs (three models $\times$ two datasets), FP8 and INT8 (weights-only) predictions were computed on an identical corpus. Pairing was verified directly from the raw per-document records: for every pair, the set of document indices in the FP8 file and the INT8 file are identical, and the underlying text at each shared index is identical (verified by hashing), confirming a fully paired design with zero unmatched documents in all 6 pairs.

Diagnostic definitions (unit of observation: document). $\text{ratio} = q_ppl/fp_ppl$ is a document-level perplexity ratio, and the reported aggregate is the mean of these document-level ratios rather than the ratio of two corpus means. kl is the document-level token-average output KL. sink is the document-level token-average D_{Sink} defined in Section 3. attn_kl and attn_w1 are summed over heads and valid query positions within each layer, summed over layers, and divided by document token count; row-level W_1 is separately divided by valid sequence length inside each row computation. Every test thus receives one value per document, although internal aggregation differs across diagnostics.

Test. Two-sided Wilcoxon signed-rank test on paired differences (FP8 − INT8) per document. Effect size is the matched-pairs rank-biserial correlation $r = (W_+ - W_-)/(W_+ + W_-) \in [-1, 1]$, with $r > 0$ meaning FP8 deviates more from the full-precision baseline than INT8 (by signed rank) and $|r| = 1$ meaning every single paired difference has the same sign. Multiple-comparison correction is Holm–Bonferroni, applied separately within each diagnostic's family of 6 tests (i.e., 5 families of $m = 6$).

Per-cell results. Table A1 reports per-cell effect size and Holm-corrected p-value for all 6 perplexity-ratio pairs, since these are individually consequential to the paper's central claim about perplexity's directional inconsistency, plus the single output-KL pair (Qwen3-4B/WikiText) that is the exception to output KL's otherwise uniform direction. Table 1 in the main text reports the aggregate significance count, direction split, and effect-size range for output KL, attention KL, attention W_1, and output D_{Sink} across all 6 pairs.

Table A1: Paired re-analysis of INT8 vs. FP8 (weights-only): perplexity ratio (all 6 pairs) and the one output-KL exception pair. Two-sided Wilcoxon signed-rank test on paired per-document differences (FP8−INT8); matched-pairs rank-biserial correlation as effect size (bounded $[-1, 1]$; positive = FP8 > INT8 by signed rank); p_{Holm} = Holm–Bonferroni-corrected p-value within each diagnostic's family of 6 tests. Output KL, attention KL, attention W_1, and output D_{Sink} are otherwise summarized in Table 1.

Diagnostic	Domain (Model)	n	Effect size r	p_{Holm}	Dir.
Perplexity ratio	IFBench (Qwen3-4B)	300	0.067	0.633	FP8>INT8
Perplexity ratio	WikiText (Qwen3-4B)	2891	0.006	0.795	FP8>INT8
Perplexity ratio	IFBench (Mistral-7B)	300	0.152	0.0675	FP8>INT8
Perplexity ratio	WikiText (Mistral-7B)	2891	−0.314	1.26×10^{-47}	✓ INT8>FP8
Perplexity ratio	IFBench (Llama-3.2-3B)	300	0.210	0.0065	✓ FP8>INT8
Perplexity ratio	WikiText (Llama-3.2-3B)	2891	0.529	9.91×10^{-133}	✓ FP8>INT8
Output KL	WikiText (Qwen3-4B)	2891	−0.201	4.47×10^{-20}	✓ INT8>FP8

C Full Downstream Analysis

Configuration-level correlation. Pooling documents from WikiText and IFBench into a single document-mean value per diagnostic and per configuration, and averaging ARC and HellaSwag equally into a single accuracy-drop and answer-flip-rate value per configuration, Table A2 reports the Spearman rank correlation between each of the five diagnostics and each downstream outcome across all fifteen configurations. All ten correlations remain significant after Holm correction across this family of tests.

Because these fifteen configurations span everything from mild weights-only INT8-versus-FP8 quantization to catastrophic activation quantization, with INT4-WO occupying an intermediate severity, this strong correlation is driven substantially by the coarse separation between mild and severe regimes rather than necessarily reflecting a fine-grained relationship. We therefore repeat the analysis restricted to the nine weights-only configurations (three models × {INT8-WO, FP8-WO, INT4-WO}). Table A3 reports the resulting correlations; note that with only nine configurations and ten correlations under test, none of these remains significant after Holm correction, so we report them as a descriptive pattern rather than an established statistical relationship.

Table A2: Spearman rank correlation between predict-side diagnostics (pooled document means across WikiText and IFBench) and downstream outcomes (ARC and HellaSwag averaged equally), across all 15 model-setting configurations. All ten correlations are Holm-significant.

Diagnostic	**vs. accuracy drop**	**vs. answer-flip rate**
Perplexity ratio	0.921	0.771
Output KL	0.771	0.943
Attention KL	0.682	0.900
Attention W_1	0.725	0.921
Output D_{Sink}	0.807	0.832

Table A3: Spearman rank correlation restricted to the nine weights-only configurations (three models $\times$ {INT8, FP8, INT4}). None of these ten correlations survives Holm correction at $n = 9$; raw p-values are shown where the correlation is discussed in the text.

Diagnostic	**vs. accuracy drop**	**vs. answer-flip rate**
Perplexity ratio	0.800 (raw $p = 0.0096$)	0.267
Output KL	0.133	0.783 (raw $p = 0.0125$)
Attention KL	0.133	0.733 (raw $p = 0.0246$)
Attention W_1	0.150	0.783 (raw $p = 0.0125$)
Output D_{Sink}	0.633	0.733 (raw $p = 0.0246$)

Within this weights-only subset, the correlation structure splits by outcome and by diagnostic in a way that does not appear in the pooled fifteen-configuration analysis. Perplexity ratio is associated with accuracy drop ($\rho = 0.800$) but not answer-flip rate ($\rho = 0.267$); output KL, attention KL, and attention W_1 are associated with answer-flip rate ($\rho = 0.73$–0.78) but not accuracy drop ($\rho = 0.13$–0.15); the Sinkhorn diagnostic falls in between, with a moderate correlation to accuracy drop ($\rho = 0.633$) alongside its correlation to flip rate. Accuracy drop is a net quantity (the number of questions that flip from correct to incorrect minus the number that flip the other way), whereas answer-flip rate counts any change of answer regardless of direction; in this weights-only regime, where net accuracy change is small and can be dominated by noise, these are evidently not interchangeable, and different diagnostics are associated with different ones.

Item-level prediction. The configuration-level correlations above raise the question of whether diagnostic magnitude at the level of an individual question predicts whether that question's answer flips. Within each of the fifteen configurations and each of the two tasks, we compare the question-level output KL, attention KL, and attention W_1 between questions whose answer flips and those whose answer does not, using a Mann–Whitney U test (3 diagnostics $\times$ 15 configurations $\times$ 2 tasks $= 90$ tests), with Holm correction across all 90.

Under a one-sided test (the pre-specified hypothesis that diagnostic magnitude is larger for flipped questions), $5/90$ tests remain significant after Holm correction; under a two-sided test, $4/90$ remain significant. In both cases, every significant test occurs under the same single quantization setting, INT8-plus-activations, and the same single task, HellaSwag, but these significant cells span all three model configurations (Qwen3-4B, Mistral-7B, and Llama-3.2-3B); none occurs on ARC, and attention KL is not significant in any test under either convention. Table A4 reports the significant cells under the one-sided convention, with the one cell that is not also significant under the two-sided convention indicated. Effect sizes, reported as the area under the curve (AUC) treating "flip" as the positive class, are modest even where significant (0.562–0.598).

Interpretation. At the level of a quantization configuration, all five diagnostics correlate strongly with downstream degradation, but restricting to the nine weights-only configurations (where, as noted above, none of the ten correlations survives Holm correction) shows a descriptive pattern in which this correlation is neither uniform across diagnostics nor across outcomes: perplexity is associated with net accuracy change, while the distributional diagnostics are associated with gross answer instability, and neither is associated with the other's preferred outcome. At the level of an

Table A4: Item-level tests that remain Holm-significant (one-sided; see text) out of 90 total tests. AUC treats "flip" as the positive class; 0.5 is chance-level separation.

Model	Setting	Task	Diagnostic	Flips	AUC	p_{Holm}
Qwen3-4B	INT8-W+A	HellaSwag	Attention W_1	427	0.598	5.34×10^{-6}
Qwen3-4B	INT8-W+A	HellaSwag	Output KL	427	0.572	0.0043
Mistral-7B	INT8-W+A	HellaSwag	Attention W_1	490	0.568	0.0083
Mistral-7B	INT8-W+A	HellaSwag	Output KL	490	0.565	0.0151
Llama-3.2-3B	INT8-W+A	HellaSwag	Attention W_1*	417	0.562	0.0321

*This cell is not significant under the two-sided convention; the remaining four cells are significant under both conventions.

individual question, diagnostic magnitude predicts an answer flip only under a single quantization setting (INT8-plus-activations) and only on one of the two tasks tested (HellaSwag), albeit across all three model configurations, with modest effect sizes even there.

D HIDDEN-STATE L^2 DETAILS

Table A5 reports, for each model, the per-setting median (over documents) of the per-document Spearman correlation between hidden-state L^2 and layer index, and the ratio of the final layer's value to the preceding layer's. Across Mistral-7B's five quantization settings (INT8-WO, FP8-WO, INT4-WO, INT8-W+A, FP8-W+A), this correlation has a median of exactly 1.000 in every setting. For Qwen3-4B, the same correlation is in the 0.988–0.994 range for four of five settings, with one exception: under INT8-plus-activations, the median correlation collapses to 0.428. For Llama-3.2-3B, the correlation is 0.9995–1.000 across all five settings. Moreover, the final layer exhibits a large discontinuity relative to the preceding layer that is consistent in direction within each model across all five quantization settings (including INT4-WO, a precision level not quantized at the activation level), but differs sharply across the three architectures both in direction and in magnitude: for Mistral-7B, the final layer is 11.1–55.9× larger than the preceding layer; for Qwen3-4B, it is instead 0.29–0.30× the preceding layer (i.e., 3–4× smaller) in the four non-exceptional settings; and for Llama-3.2-3B, it is 1.65–2.07× larger, a distinct third pattern that is neither the sharp Mistral-style amplification nor the Qwen-style attenuation. That this discontinuity is stable across weight-only precision levels (INT8, FP8, and INT4 all show the same pattern within a given model) while differing in both sign and magnitude across the three model families we tested is evidence that it is model-family-dependent and not attributable solely to quantization precision – but neither its direction nor its magnitude can be assumed to generalize to untested model families, and we do not identify architecture as the confirmed cause since the three model families tested also differ in other respects.

Table A5: Hidden-state L^2 depth dependence, by model. For each quantization setting, we compute the per-document Spearman correlation between hidden-state L^2 and layer index and take its median across documents; "Spearman (layer index)" reports the value (or range of values) of this per-setting median across the settings listed for that model in the "Settings" column. "Final/preceding ratio" is the final layer's value divided by the preceding layer's.

Model	Settings	Spearman (layer index)	Final/preceding ratio
Mistral-7B	all 5 settings	1.000	11.1–55.9×
Qwen3-4B	4 of 5 (excl. INT8-W+A)	0.988–0.994	0.29–0.30×
Qwen3-4B	INT8-W+A (exception)	0.428	–
Llama-3.2-3B	all 5 settings	0.9995–1.000	1.65–2.07×

E Extended Limitations and Future Directions

Fixed-prefix evaluation and the gap to path-law comparison. Our main experiments evaluate fixed-prefix conditional distributions; free-running autoregressive generations remain outside this comparison. This is intentional: it isolates local, quantization-induced deformation without confounding it with the history divergence that free-running generation introduces once sampled tokens differ. Once $x^{\mathrm{fp}}_{<t} \neq x^{q}_{<t}$, the subsequent conditional distributions are conditioned on different histories, so a comparison at that point cannot separate the local deformation quantization induces from the accumulated consequence of the two models having diverged; this is why we use a same-prefix design in the main experiments. Evaluating the full free-running case instead means comparing the path measures $P^{(T)}_{\mathrm{fp}}$ and $P^{(T)}_{q}$ that M_{fp} and M_q induce over complete output trajectories, which concerns full trajectory distributions and cannot be inferred from any single realized pair; turning trajectory pairs into such a statement requires choosing how the two processes are coupled, aggregating over enough trajectories for that coupling to be meaningful, and respecting the autoregressive history-dependence and temporal filtration of the process without treating trajectories as exchangeable static objects. Separately, we also explored a direct token-embedding distance between free-running generated sequences (Section 3); as an empirical, methodological observation independent of the path-law argument above, this simple trajectory-level distance did not behave as a stable diagnostic in our exploration, which is part of why we turn to the same-prefix design. Our fixed-prefix diagnostics are local conditional objects cut from this same underlying process; they do not by themselves evaluate the trajectories that the quantized model would produce after repeatedly conditioning on its own outputs.

Asymmetry of the local diagnostic and the case for adapted comparison. Section 3 showed that single-step conditional KL is an exact component of path-measure KL only when the outer expectation is taken over prefixes drawn from M_{fp}'s own rollout distribution, and that our reported document-level statistic instead averages over ground-truth dataset prefixes – a related but distinct quantity. Even under the rollout-expectation reading, the resulting equality is asymmetric: it measures how M_q differs from M_{fp} on prefixes typical under M_{fp}, but not how M_q behaves on prefixes that become likely only after its own errors accumulate. A symmetric comparison addresses the directionality of discrepancy, whereas a causally adapted comparison constrains couplings to respect temporal information flow.

Toward adapted Wasserstein and path-space diagnostics. Both limitations above point to the same natural next step: comparing the full joint distributions $P^{(T)}_{\mathrm{fp}}$ and $P^{(T)}_{q}$ that M_{fp} and M_q induce over complete output trajectories, rather than a collection of local conditional slices of them. Ordinary optimal transport on sequence space can incorporate path-dependent costs, but it does not by itself enforce nonanticipative coupling constraints that respect the temporal filtration of autoregressive generation. Candidates that impose such constraints include nested distances Pflug & Pichler (2012) and causal or bicausal transport formulations Backhoff et al. (2017), which restrict the coupling between $P^{(T)}_{\mathrm{fp}}$ and $P^{(T)}_{q}$ to respect the order in which information becomes available to each process, and may capture exposure-bias-like compounding errors that single-step, same-prefix diagnostics cannot. We have not implemented or evaluated these path-space formulations in the present study; we therefore present them as natural candidates for future work rather than as the uniquely appropriate destination.

Downstream task performance. Section 4.7 evaluates two multiple-choice benchmarks, ARC-Challenge and HellaSwag, which are themselves scored via prompt-conditioned choice likelihoods rather than free-running generation; we do not claim that the proposed diagnostics are a general surrogate predictor of downstream accuracy, and the evidence in Section 4.7 directly supports this: configuration-level correlation is strong, but it does not survive restriction to the weights-only regime as a statistically established relationship, and item-level prediction of individual answer flips is largely absent outside a single quantization setting (INT8-plus-activations) and a single task (HellaSwag). What remains genuinely untested is task accuracy under free-running, open-ended generation, where answer extraction and scoring operate on the model's own sampled outputs, with no fixed set of candidate continuations. The natural measure-level extension of our fixed-prefix diagnostics to that setting is again the full path measure discussed above in place of a per-token or per-choice

functional; establishing a principled connection from local conditional deformation to open-ended generation quality requires the path-space and adapted-transport tools just described, together with decoding-aware stability analysis and task-specific functional bounds. We leave that connection to future work while isolating, in the present study, the local measure-valued perturbations captured by fixed-prefix conditional distributions and attention measures, together with their empirically measured (and empirically limited) relationship to multiple-choice downstream accuracy.

Geometry for output transport. The limited additional gain of the output Sinkhorn diagnostic may reflect a limitation of the underlying embedding geometry rather than of optimal transport. Unlike attention positions, whose one-dimensional geometry has a direct operational interpretation, vocabulary embeddings are learned internal coordinates entangling semantic, syntactic, frequency, tokenization, and architectural effects; a geometry designed to capture semantic proximity specifically, in contrast to one that inherits those entangled factors from native embeddings, could change the relative diagnostic value of output transport, and the present experiments do not rule out that possibility. Designing geometries better suited to probability-measure evaluation is an important direction for future work. This includes output-head geometry, context-dependent token representations, Fisher-like local geometries, and learned semantic geometries calibrated against human preference or downstream task degradation.

Aggregation and statistical discrimination. All inferential tests use documents, rather than individual heads, layers, or token positions, as the unit of observation. A full variance decomposition across tokens, heads, and layers remains an open question.

Broader model and quantizer coverage. Our study covers three model families, two datasets, and five quantization configurations (INT8, FP8, and INT4 weights-only, and INT8 and FP8 weights-plus-activations). Broader corpus coverage was limited by available computational resources; C4 and FineWeb were therefore excluded from the main analysis rather than evaluated with substantially smaller and uneven sample sizes. Extending the main audit to these corpora, once sufficient computational resources are available, is a direct next step using the same method. The study does not yet include additional model families, lower-bit weights-only regimes below INT4, or calibrated mitigation schemes. Broader evaluation across architectures, bit widths, quantization algorithms, and calibration procedures is necessary before making universal claims about precision formats or diagnostics; the architecture-dependence already visible across three models in the layerwise and hidden-state results above suggests that such universal claims should be treated with particular caution. Our current experiments use naive weights-only and weights-and-activations PTQ; they do not apply methods such as SmoothQuant Xiao et al. (2023), AWQ Lin et al. (2024), ZeroQuant Yao et al. (2022), or OmniQuant Shao et al. (2024). Whether the orderings and exceptions reported here persist after such mitigations is an open question that the diagnostic framework is well-suited to test, but does not itself answer.